User Manual for

# TASSEL

***Trait Analysis by aSSociation, Evolution and Linkage***

Version 5.0

The Buckler Lab at Cornell University

(July 17, 2014)

www.maizegenetics.net/tassel

**Disclaimer**: While the Buckler Lab at Cornell University has performed extensive testing and results are, in general, reliable, correct or appropriate. Results are not guaranteed for any specific set of data. It is strongly recommended that users validate TASSEL results with other software.

**Further help**: Additional help is available beyond this document. Users are welcome to report bugs, request new features through the TASSEL website. Questions are also welcome to our current team members. For more quick and precise answers, please address your questions to the most pertinent person:


| | |
|---|---|
| **Tassel User Group (recommended)** | **http://groups.google.com/group/tassel** tassel@googlegroups.com |
| **General Information** | **Ed Buckler (Project leader)** esb33@cornell.edu |
| **Data Import, Pipeline** | **Terry Casstevens** tmc46@cornell.edu |
| **Statistical Analysis** | **Peter Bradbury** pjb39@cornell.edu |


**Contributors**: Ed Buckler, Terry Casstevens, Peter Bradbury, Zhiwu Zhang, Dallas Kroon, Jeff Glaubitz, Kelly Swarts, Jason Wallace, Fei Lu, Alberto Romero, Cinta Romay, Eli Rodgers-Melnick, Alexander Lipka, Sara Miller, James Harriman, Yogesh Ramdoss, Michael Oak, Karin Holmberg, Natalie Stevens, and Yang Zhang.

**Citations**:

Overall Package:
Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Genotyping by Sequencing:
Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE* **9**(2): e90346

Mixed Model GWAS:
Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**:355-360.

The TASSEL project is supported by the National Science Foundation and the USDA-ARS.

**Reference Links**:

Main Web Site: http://www.maizegenetics.net/tassel
Open source code: https://bitbucket.org/tasseladmin/tassel-5-source
Wiki: https://bitbucket.org/tasseladmin/tassel-5-source/wiki

## Table of Contents


- Introduction
- Getting Started
  - Executing TASSEL
  - Open Source Code
  - Software Development Tools
  - Graphical Interface
  - Pipeline (Command Line Interface)
  - GBS Pipeline
- File Menu
    - Save Data Tree
    - Open Data Tree
    - Save Data Tree As…
    - Open Data Tree…
    - Set Preferences
- Data Menu
  - Load
    - Hapmap
    - HDF5 (Hierarchical Data Format version 5)
    - VCF (Variant Call Format)
    - Plink
    - Projection Alignment
    - Phylip
    - FASTA
    - Numerical Data
      - Trait format
      - Covariate Format
      - Marker Values as Numerical Co-variates
    - Square Numerical Matrix
    - Table Report
    - TOPM (Tags on Physical Map)
  - Export
  - Sort Genotype File
  - Transform
    - Genotype Numericalization
      - Collapse Non Major Alleles
      - Separate Alleles
    - Transform and/or Standardize Data
    - Impute Phenotype
    - PCA
  - Synonymizer (Synonymize Taxa Names)
  - Intersect Join
    - Command
  - Union Join
    - Command
  - Merge Genotype Tables
    - Command

Notes
Separate
Homozygous Genotype
Impute Menu
Genotypic Imputation
Filter Menu
Sites
Site Names
Taxa Names
Taxa
Traits
Analysis Menu
Diversity
Linkage Disequilibrium
Cladogram
Kinship
GLM (General Linear Model)
MLM (Mixed Linear Model)
Genomic Selection (using Ridge Regression)
Geno Summary
Stepwise
Results Menu
Table
Archaeopteryx Tree
2D Plot
LD Plot
Chart
QQ Plot
Manhattan Plot
GBS Menu
Help Menu
Help Manual
About
Show Memory
Logging
Tutorial
Missing Phenotype Imputation
Principal Component Analysis
Estimation of Kinship using genetic markers
Association analysis using GLM
Association analysis using MLM
Appendix
Nucleotide Codes (Derived from IUPAC)
TASSEL Tutorial Data sets
Frequently Asked Questions
REFERENCES

# Introduction

While TASSEL has changed considerably since its initial public release in 2001, its primary function continues to b providing tools to investigate the relationship between phenotypes and genotypes[1]. TASSEL has functionality fo association study, evaluating evolutionary relationships, analysis of linkage disequilibrium, principal component analy cluster analysis, missing data imputation and data visualization. TASSEL development has been led by a group focuse on maize genetics and genomics, and for these reasons that software has design and computational optimizations tha account for the biology found in many plants and breeding situations. Compared to human genetics, many crops are highly diverse both at the nucleotide level and structural variations (10-50X greater than humans), inbreeding i common, large families are common, and whole genome prediction is being applied daily to real world problems. Thes biological differences lead to some different optimizations that are of use to many biological systems outside of crops.

One of the design elements driving TASSEL development has been the need to analyze ever larger sets of data[2]. TASSEL5 has at its heart lots of design optimizations for big data, including:

- Bit level encoding of nucleotides so genetic distance and linkage disequilibrium estimates can be made very quickly (20-50X speed increases).
- Extensive use the HDF5 file format, which has been developed as a robust element of many climate modelers for matrix style data
- Tools for extracting and calling SNPs from extensive Genotyping-by-Sequencing data (tested for 60,000 samples by over 2.5 million SNPs and 96 million sequence alleles).
- Projection and imputation procedures that are optimized for the large families in crops. Some of these optimizations permit memory and computational improvements of >100,000 fold.
- Mixed models based on DNA relationships have come to dominate GWP (Meuwissen et al 2001) and GWAS (Yu et al 2006), yet these models can be slow to solve. TASSEL has been a test bed and implements some of the most best optimizations, such as EMMA (Kang at al 2008), plus approaches optimize variance components once P3D (Zhang et al 2010) and EMMAX (Kang et al 2010). Compression algorithms are also available (Zhang et al 2010). When used correctly, these optimizations make powerful GWAS computationally possible.
- The code is being continually optimized for larger numbers of cores and clusters. For example, we generally run imputation on 64-core machines. And while Java provides some excellent is interoperability between systems, its code is about 2-fold slower than optimized C libraries, and 10-fold slower than GPU processing for some problems. TASSEL5 is building out connection layers directly to native code, when these efficiencies are need.

TASSEL was designed for a wide range of users, including those not expert in statistical genetics or computer science A GWAS using the mixed linear model method to incorporate information about population structure[6-8] and cryptic relationships[9] can be performed by in a few steps by "clicking" on the proper choices using a graphic interface. All th processes necessary for the analysis are performed automatically, including importing phenotypic and genotype data imputing missing data (phenotype or genotype), filtering markers on minor allele frequency, generating princip components and a kinship matrix to represent population structure and cryptic relationships, optimizing compressio level and performing GWAS.

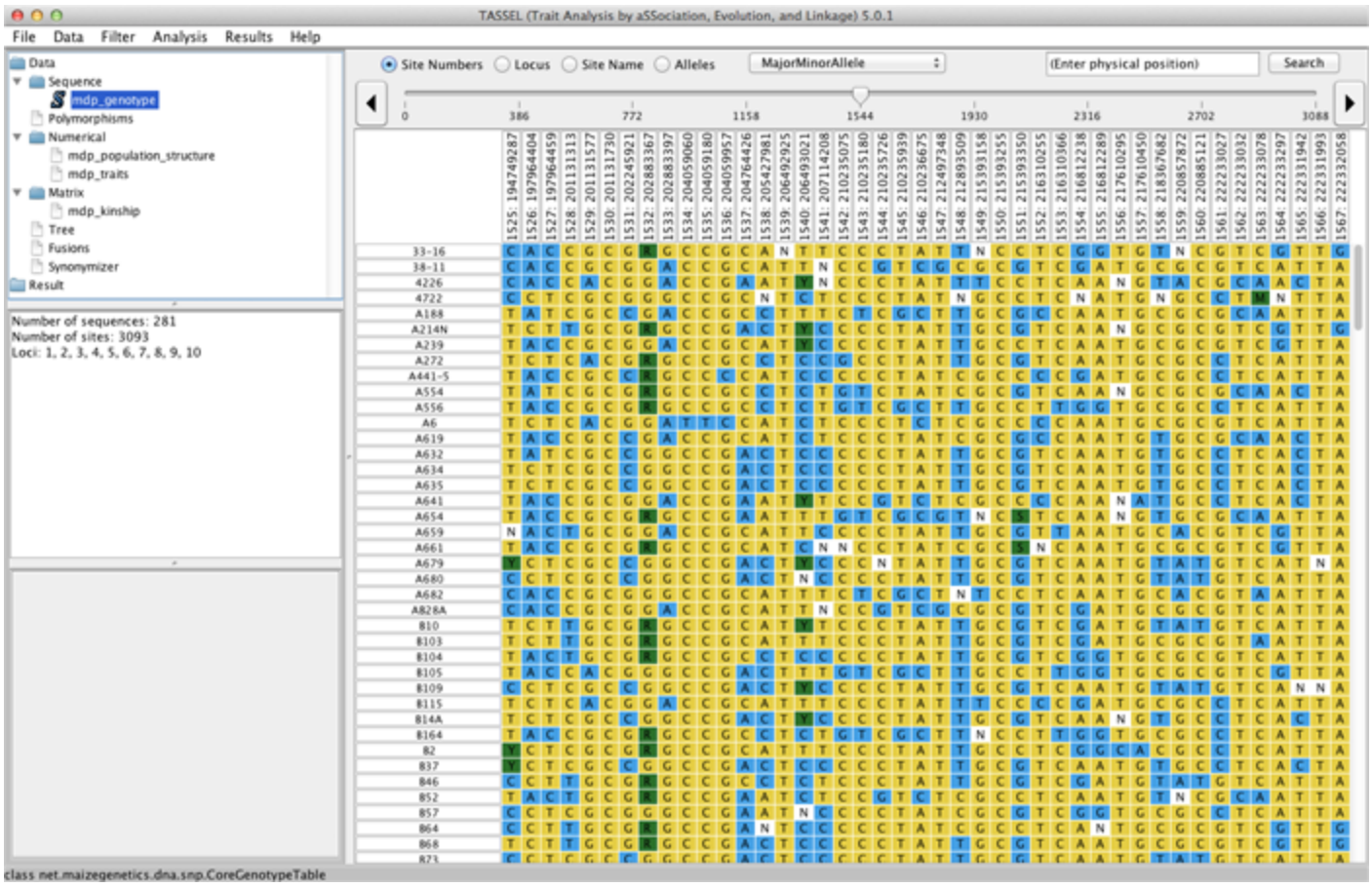
The command-line version of TASSEL, called the Pipeline, provides users the ability to program tasks using a script instead of the graphic user interface (GUI). This feature allows researchers to define tasks using a few lines of code and provides the ability to use TASSEL as part of an analysis pipeline or to perform simulation studies. We are also buildin a larger community of scientist developers that are adding functionality to this platform and working together to improv

the system.  So throughout this user manual you will see how to do most things three different ways - with the GUI, with the pipeline, and with the API (application programming interface).

TASSEL is written in Java, thereby enabling its use with virtually any operating system. It can be installed using Java Web Start technology by simply clicking on a link at www.maizegenetics.net/tassel. A stand-alone version of TASSEL can also be downloaded to use in pipeline mode or in any situation where the user wishes to start the software from a command line.

# Getting Started

A quick way to get started using TASSEL is to load the tutorial data and try performing analyses. However, because some of the necessary steps may not be intuitive, we recommend that new users follow the tutorial at end of this manual. The objective of this section is to provide information necessary to install and start TASSEL software and to provide a brief overview of the interface.




## 1.1 Executing TASSEL

http://www.maizegenetics.net/tassel/docs/ExecutingTassel.pdf

## 1.2 Open Source Code

Open source code for TASSEL is available at: https://bitbucket.org/tasseladmin/tassel-5-source. The package uses a number of other libraries that are included in the TASSEL distribution. These include a modified version of the PAL library (http://www.cebl.auckland.ac.nz/pal-project/), the COLT library (http://dsd.lbl.gov/~hoschek/colt/), jFreeChart (http://www.jfree.org/jfreechart/), Guava (Google Core Libraries) (https://code.google.com/p/guava-libraries), JUnit (http://junit.org), Archaeopteryx (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx), and BioJava (http://www.biojava.org).

## 1.3 Software Development Tools

jProfiler (http://www.ej-technologies.com/products/jprofiler/overview.html)
install4j (http://www.ej-technologies.com/products/install4j/overview.html)
NetBeans IDE (https://netbeans.org)
Eclipse (http://www.eclipse.org)
IntelliJ (http://www.jetbrains.com/idea)
Structure101 (http://structure101.com)
TeamViewer (http://www.teamviewer.com)
Bitbucket (https://bitbucket.org)
sourceforge (http://sourceforge.net)
JIRA (https://www.atlassian.com/software/jira)
Tower (http://www.git-tower.com)

## 1.4 Graphical Interface

TASSEL is organized into five main panels. 1) At the top menus control functions. 2) The Data Tree at the top left organizes data sets and results. Data set(s) displayed in the Data Tree must first be selected before a desired function or analysis can be performed. To select multiple data sets, press the CTRL (or Command for Mac) key while selecting the data sets. 3) The Report Panel is located below the Data Tree. It displays information about a selected data set from the Data Tree, such as the type of data and how it was created. 4) The Progress Monitoring Panel below the Report Panel shows the progress of running tasks and has buttons that can cancel tasks. 5) The Main Panel occupies the right side of the viewing area, and displays the content of the selected data set from the Data Tree.

## 1.5 Pipeline (Command Line Interface)

http://www.maizegenetics.net/tassel/docs/TasselPipelineCLI.pdf

## 1.6 GBS Pipeline

http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf

# 2 File Menu

The data tree can be saved in a binary format.

### 2.1.1 Save Data Tree

This feature allows you to save the entire contents of the Data Tree panel to a default location. This is helpful when th user does not wish to recreate a Data Tree panel that is already well populated with information the next time they initializes the program. To save a Data Tree, select **File > Save Data Tree**.

### 2.1.2 Open Data Tree

To restore a Data Tree that was saved previously saved, select **File > Open Data Tree**.

### 2.1.3 Save Data Tree As…

To save the Data Tree to a specific location or to give it a specific name, select **File > Save Data Tree As….**

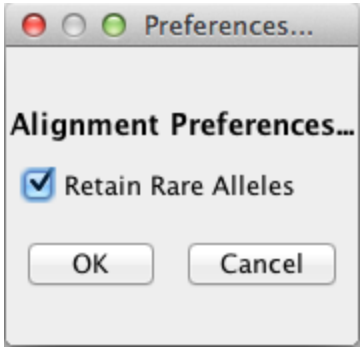
### 2.1.4 Open Data Tree…

To restore a Data Tree from a specific location, select **File** > **Open Data Tree…**

**NOTE: The information outlined above for saving a Data Tree is applicable to files that are, in general, version specific. When a new version of TASSEL is released, a data tree saved with a previous version might not load to the version. For longer term storage, the best practice is to save individual data sets rather than the entire data tree.**

### 2.1.5 Set Preferences

Currently there is only one preference. That is whether to retain "rare" alleles. This is irrelevant for nucleotide data (A, C, G, T, -, +, N) because at that number of states, there is no data lost. Potentially with other types of data, it could exceed the 14 max (per site) number of allele states. If you "Retain Rare Alleles", the lower frequency allele values wil be consolidated into a rare (Z) state. Otherwise, those lower frequency alleles are changed to unknown (N).




# 3 Data Menu

The Data Menu has options to import and export data sets, as well as, other data manipulate functions.

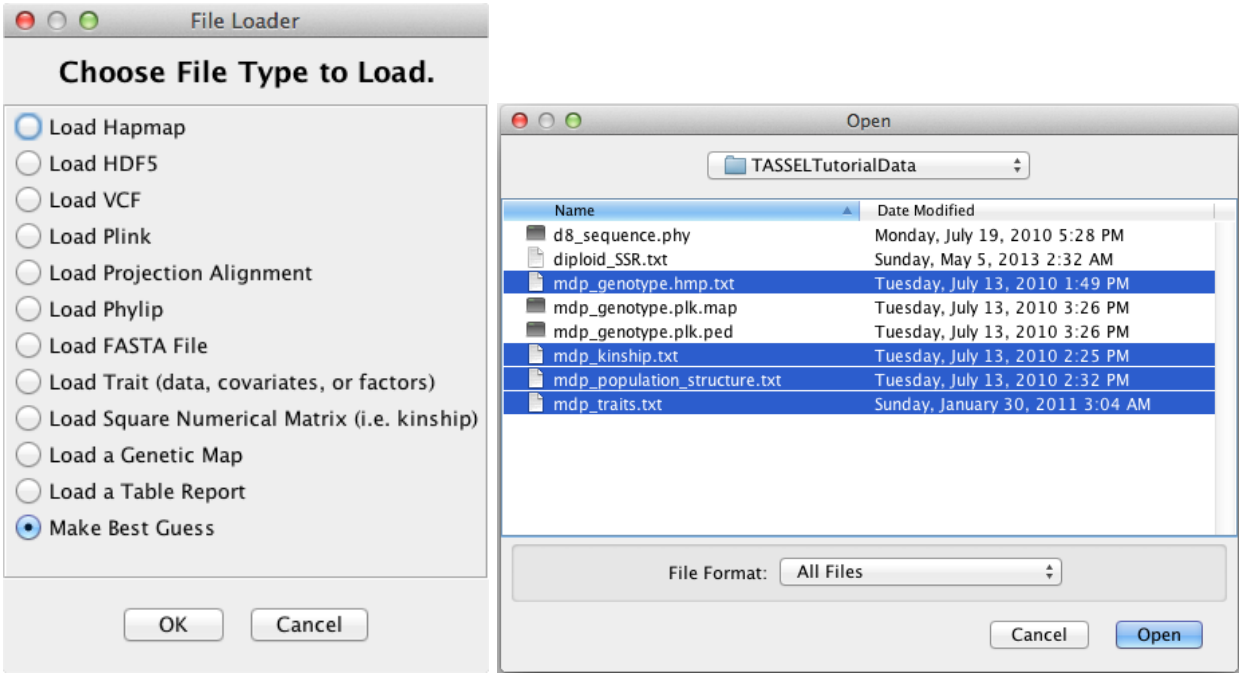
## 3.1 Load

Load provides options to import files for genotypes, phenotypes, populations structure, and kinship matrices, etc.

The tutorial data can be downloaded from the TASSEL website at this link:

http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip.

To use the data, the zip file must be uncompressed and saved on your local machine. These tutorial files will loa correctly with the "Make Best Guess" option. Multiple files can be imported simultaneously by highlighting them fi (holding Shift or Control key while clicking) and then clicking the Open button.




### 3.1.1 Hapmap

Hapmap is a text based file format for storing sequence data. All the information for a series of SNPs as well as the germplasm lines are stored in one file. The first row contains the header labels, and each additional row contains all th information associated with a single SNP. The first 11 columns describe attributes of the SNP, while the following columns describe the SNP value for a single germplasm line. The first 12 columns of the first row should look like thi where "Line 1" is the beginning of germplasm line names.

| rs# | alleles | chrom | pos | strand | assembly# | center | protLSID | assayLSID | panelLSID | QCcode | Line 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|

While all 11 header columns are required, not all 11 of the columns need to be filled in for TASSEL to correctly interpret the data. The only required fields are “chrom”, Chromosome name, and “pos”, Position. In the example below, genotype values are represented by 2 characters (i.e. AA). Note that you can record those as single character values (see “Nucleotide Codes” in the Appendix).

For TASSEL to correctly read Hapmap data, the data must be in order of position within each chromosome, and the file should be TAB delimited (example below is in Excel only for easy viewing). If some of the data is missing the corre number of TABs must still be present, so that TASSEL can properly assign data to columns.

| rs# | alleles | chrom | pos | strand | assembly# | center | protLSID | assayLSID | panel | QCcode | 33-16 | 38-11 | 4226 | 4722 | A188 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PZB00859.1 | A/C | 1 | 157104 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | CC | AA |
| PZA01271.1 | C/G | 1 | 1947984 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | GG | CC | GG | CC |
| PZA03613.2 | G/T | 1 | 2914066 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | GG | GG | GG | GG | GG |
| PZA03613.1 | A/T | 1 | 2914171 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | TT | TT | TT | TT |
| PZA03614.2 | A/G | 1 | 2915078 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | GG | GG | GG | GG | GG |
| PZA03614.1 | A/T | 1 | 2915242 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | TT | TT | TT | TT |
| PZA00258.3 | C/G | 1 | 2973508 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | GG | CC | CC | CG | CC |
| PZA02962.13 | A/T | 1 | 3205252 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | TT | TT | TT | TT |
| PZA02962.14 | C/G | 1 | 3205262 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | CC | CC |
| PZA00599.25 | C/T | 1 | 3206090 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | TT | CC | TT | TT |
| PZA02129.1 | C/T | 1 | 3706018 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | CC | CC | CC | CC |
| PZA00393.1 | C/T | 1 | 4175293 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | TT | TT | CC | TT |
| PZA02869.8 | C/T | 1 | 4429897 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | TT | CC | NN | CC |
| PZA02869.4 | C/G | 1 | 4429927 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | NN | GG |
| PZA02869.2 | C/T | 1 | 4430055 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | NN | TT | TT | CC | TT |
| PZA02032.1 | A/T | 1 | 4490461 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | AA | TT | AA | AA | AA |
| zagl1.5 | A/T | 1 | 4835434 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | AA | NN | AA | AA | AA |
| zagl1.2 | A/C | 1 | 4835558 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | CC | CC |
| zagl1.6 | C/T | 1 | 4835658 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | TT | TT | TT | TT | TT |
| PZD00081.2 | C/T | 1 | 4836542 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | CC | CC |
| zagl1.1 | A/C | 1 | 4912526 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | AA | AA | AA | AA | AA |
| PZB00919.1 | A/C | 1 | 5353319 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | CC | CC | CC | CC | AA |
| PZB00919.2 | G/T | 1 | 5353655 | + | AGPv1 | Panzea | NA | NA | maize282 | NA | GG | GG | GG | GG | GG |

### 3.1.2 HDF5 (Hierarchical Data Format version 5)

http://www.hdfgroup.org/HDF5

### 3.1.3 VCF (Variant Call Format)

http://www.1000genomes.org/wiki/analysis/variant-call-format/vcf-variant-call-format-version-42

### 3.1.4 Plink

Plink is a whole genome association analysis tool set, which comes with its own text based data format. The data is stored in a set of two files, a .map file and a .ped file.

The .ped file contains all the SNP values and has six mandatory header columns for Family ID, Individual ID, Paternal ID, Maternal ID, Sex and Phenotype. TASSEL only requires that the Individual ID field be filled in. Each row of the .ped file describes a single germplasm line. Notice in Plink, an unknown character is represented with a '0'. However in TASSEL an unknown character is represented with a 'N', and '0' is used to represent heterozygous indel. TASSEL will automatically convert between the '0' and the 'N'. Any exported Plink files will represent the heterozygous indel with a

'+' (insertion) and a '-' (deletion).

The .map file describes all the SNPs in the associated .ped file, where each row provides information on one SNP. The .map file must contain exactly four columns: Chromosome, rs#, Genetic distance and Position. TASSEL does not require the Genetic distance field to be filled in.

Both files should be TAB delimited.

For a more detailed description on the data format, please visit the Plink basic usage and data formats webpage: (http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml).

### 3.1.5 Projection Alignment

### 3.1.6 Phylip

Details on Phylip format are described at the following website:
http://evolution.genetics.washington.edu/phylip/doc/sequence.html

### 3.1.7 FASTA

### 3.1.8 Numerical Data

This type of format is used for trait and covariate data such as population structure. Similar to sequence alignmen genotype data, numerical data also consists of two parts: a header that defines data structure and a body containing the main data. Tabs should be used as delimiters. However, any white space character such as blank will be treated as a delimiter as well. As a result, embedded blanks in names will cause data to be imported incorrectly. We suggest representing missing values using “NA”, or “NaN”. However, any text value (e.g. “?”) will be interpreted as missin data. There are several formats for numerical data to fit the requirement for modeling. Trait data (dependent variables) can be imported by starting the first line with “<Trait>” and following that with the trait names. Additional classifiers m also be included in subsequent header rows by starting the row with “<Header name=xxx>” followed by a name for each column of data. For instance, to define environments, start the second header row with “<Header name=env>”.

Comment lines may be inserted at the beginning of the file. Comment line begins with the character “#”.

#### 3.1.8.1 Trait format

This format does not require users to provide information on number of rows and columns. The file starts with the key word <Trait> followed by names of columns. The column for line should not be labeled.

Example 1, simple list of trait values:

```
<Trait>      EarHT dpoll EarDia
811    59.5  NA    NA
33-16  64.75 64.5  NA
```

```
38-11 92.25 68.5  37.897
4226  65.5  59.5  32.21933
4722  81.13 71.5  32.421
A188  27.5  62    31.419
…
```

Example 2, traits data collected in multiple environments:

```
<Trait>     EarHT PlantHT     EarHT PlantHt
<Header name=env> Loc1  Loc1  Loc2  Loc2
811   59.5  NA    NA    NA
33-16 64.75 121.5 NA    NA
38-11 92.25 153.8 37.897      83.4
4226  65.5  130.1 32.21933    82.1
4722  81.13 165.7 32.421      90.1
A188  27.5  110.2 31.419      79.6
…
```

#### 3.1.8.2 Covariate Format

Covariate data uses the same format as trait data except that the first line must be "<Covariate>". This line tells TASSEL that the variables in this file will be used as covariates not as dependent variables. This is the format to use fo population structure covariates.

```
<Covariate>
<Trait>     Q1    Q2    Q3
33-16 0.014 0.972 0.014
38-11 0.003 0.993 0.004
4226  0.071 0.917 0.012
4722  0.035 0.854 0.111
A188  0.013 0.982 0.005
…
```

#### 3.1.8.3 Marker Values as Numerical Co-variates

In some cases, a user may wish to have marker values treated as numerical co-variates. If the first line of the file is "<Numeric>", then the data will be imported as numeric data but used as marker data in GLM and MLM.

```
<Numeric>
<Marker> m1 m2 m3 m4 m5
33-16 0 1 1 0 0
38-11 0 0 1 0.3 0
4226 0 1 1 0.5 0
```

### 3.1.9 Square Numerical Matrix

Kinship can be calculated externally from pedigrees by using SAS Proc Inbreeding[18] or from markers by using one of several available software packages. The following format is provided to import the resulting kinship estimates:

If n represents the number of taxa, the format for kinship files is as follows:

**n**
**Taxa1Name r11 r12 … r1n**
**Taxa2Name r21 r22 … r2n**
…
**TaxanName rn1 rn2 … rnn**

Here rij (i, j=1,2, …, n) is the element in the kinship matrix located at row i and column j.

Missing values are not allowed for kinship matrix.

**Important note:** The current format is different from the format used in TASSEL version 2.0 or lower.

### 3.1.10 Table Report

Data can be imported as tab delimited text files. The first row of the file will be interpreted as column labels and the remaining rows as rows in the table.

### 3.1.11 TOPM (Tags on Physical Map)

## *3.2 Export*

Options are provided to export sequence data: Hapmap, Plink, Phylip (Sequential or Interleaved). Phenotypes and covariate data is exported as numerical trait data. Table Reports are exported as a tab delimited table. For numerical data, the function of Export is similar to the Table function in Results mode.

## 3.3 Sort Genotype File

TASSEL5 has strict requirements for the sites in a genotype file. Each site must be unique (as defined by its locus/chromosome, position, and name) and they must be in order in the file. Genotype files produced by other programs (and also earlier versions of TASSEL) often do not meet this second requirement and throw an error when TASSEL tries to load them. It can be difficult to recreate TASSEL's internal sort order by hand, so this plugin allows the user to sort an input genotype file according to TASSEL's rules and output it to a new file ready for further analysis. (This sort is not done automatically at load time because the computational cost for sorting large files can be very large. We feel it's better for users to know what they're getting into instead of being surprised by it.) There is currently only support for sorting Hapmap and VCF files.

To sort a genotype file from the GUI, just select Data -> Sort Genotype File and fill in the appropriate parameters in the popup dialog.

To sort a file from the command line, use the following command:

```
run_pipeline.pl -SortGenotypeTablePlugin -inputFile [filename] -outputFile
[filename] -fileType [Hapmap or VCF]
```

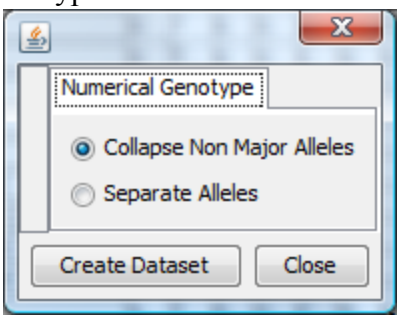
The -fileType flag is optional and is only needed if the input file's extension doesn't match a known file extensio (".hmp.txt", ".vcf", etc.).

## 3.4 Transform

This suite of functions allows multiple data manipulation on genotype and phenotype (numerical) data. When a genotyp data set is selected, the data are transformed to numbers. When a numerical data set is selected, mathematical transformation, data imputation and principal component analysis (PCA) can be performed. The Transform columns tags will be displayed in a Data dialog box with three tabs: Trans, Impute and PCA.

### 3.4.1 Genotype Numericalization

Two options are provided to transform genotype from character to numerical as shown in the following dialog box.




#### 3.4.1.1 Collapse Non Major Alleles

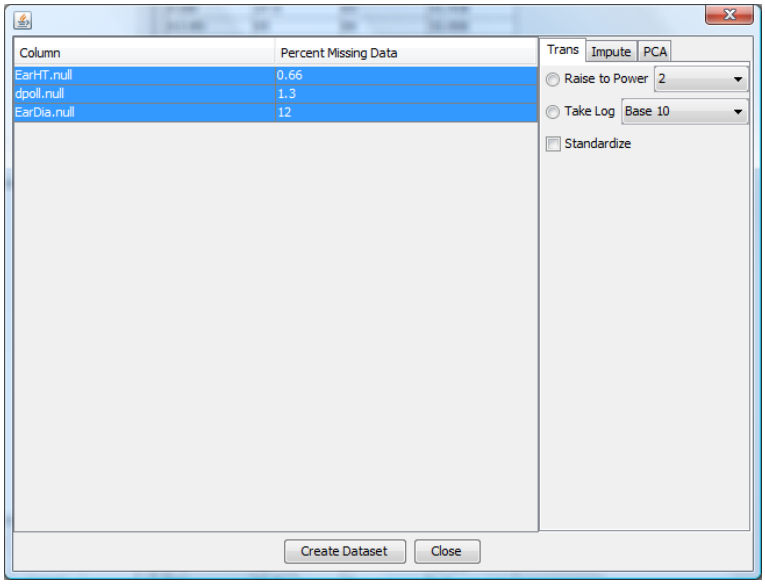
This function assigns 1 to the major allele and 0 to any other alleles. The converted genotypes are saved in a new numerical data set.

#### 3.4.1.2 Separate Alleles

This function assigns an indicator (1 for present and 0 for absent) for each allele. The converted genotypes are saved in a new numerical data set.

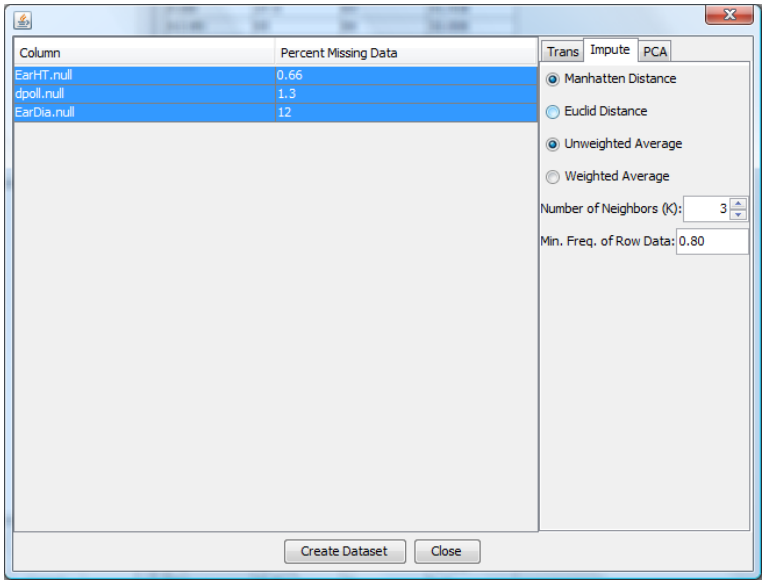
### 3.4.2 Transform and/or Standardize Data

The **Trans** dialog box is the default selection, as shown below. In the **Column** list, select the column(s) you wish to transform. Then select the type of transformation you wish to execute. Selecting the **Standardize** checkbox will transform data by subtracting the column mean from the value of the trait and then dividing by the column's standard deviation. Clicking on the **Create Data set** button will result in the placement of a dataset containing only the selected columns in the Data Tree.




### 3.4.3 Impute Phenotype

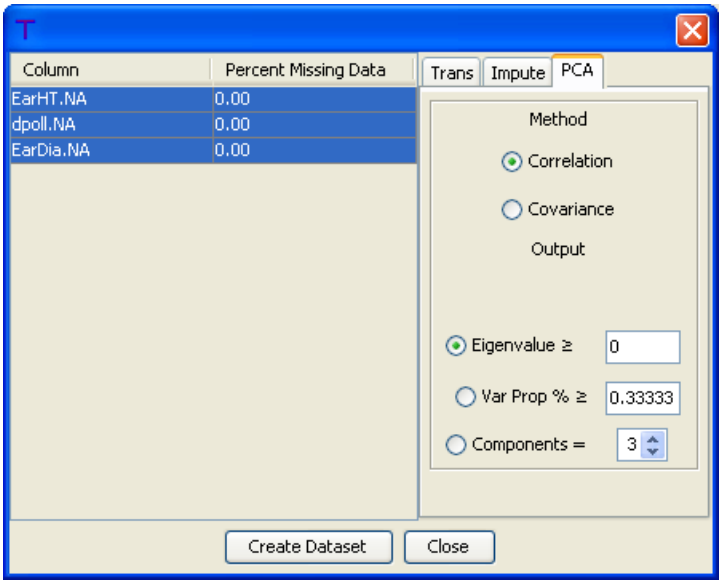
The k-nearest-neighbor algorithm[20] is used to impute missing phenotype data. If data is missing for a taxon for one o the traits, the algorithm finds other taxa (neighbors) that are most like it for the non-missing traits. It uses the average o the neighbors to impute the missing data. Click on the **Impute** tab to display the following:




### 3.4.4 PCA

Principal component analysis (PCA) can only be performed on a numerical data set without missing values. Two methods are available: correlation or covariance. This determines whether a correlation or covariance matrix will be used as the basis for the analysis. The default, correlation, is a reasonable choice for genetic data. The number of PCA axes in the output data set can be controlled by selecting either of the minimum eigen value associated with each axis

the minimum percent of the variance captured by an axis or the number of axes. The resulting axes will be sorted by th amount of variance each captures.

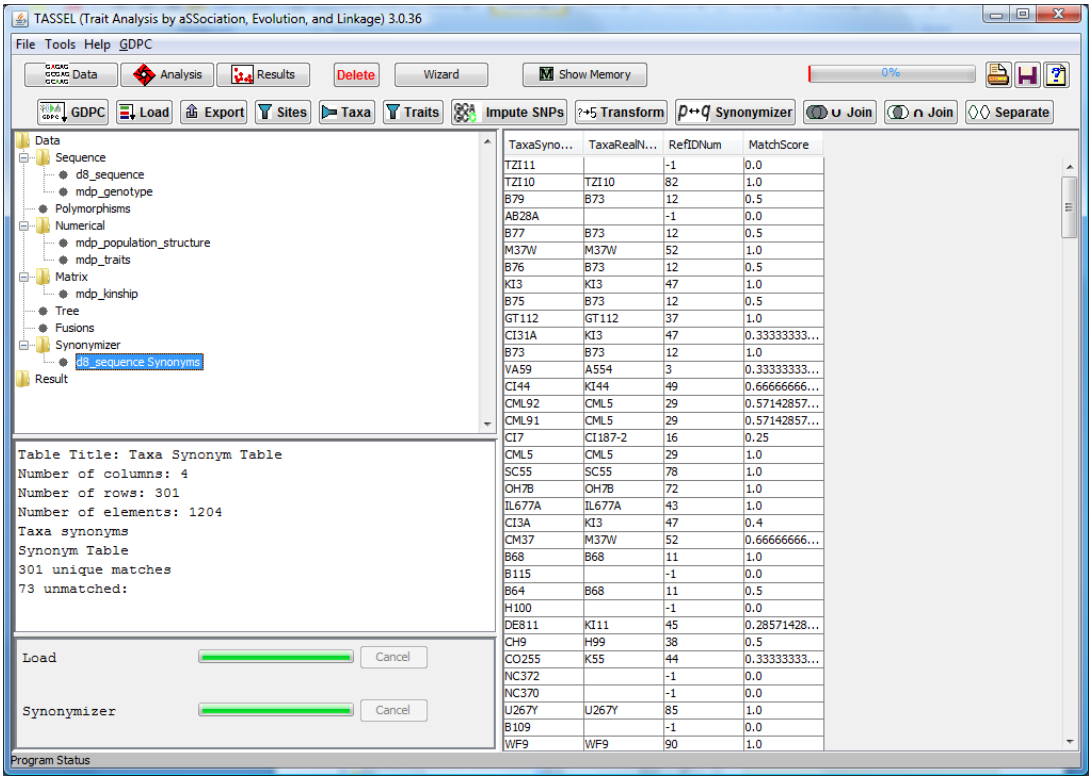



## *3.5 Synonymizer (Synonymize Taxa Names)*

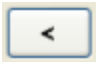
This button makes taxa names uniform to permit the joining of data sets.

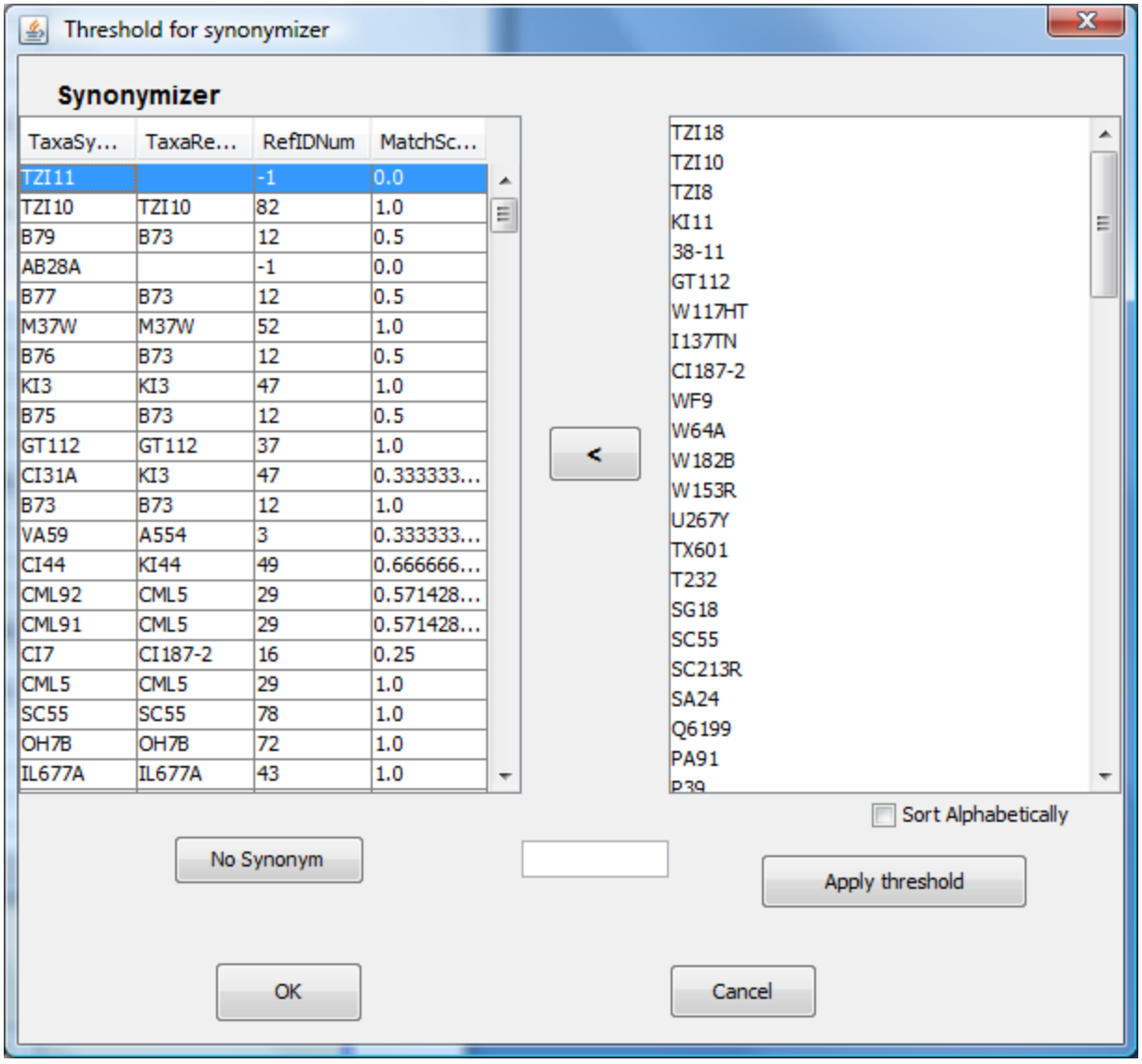
The join functions that generate fused data sets work by matching taxa names. Consequently, if multiple names exist fo a given taxon (an added suffix, alternative spellings, different naming conventions, etc.) then the two data sets will n join correctly. To help remedy this, the Synonymizer function allows the taxa names of one data set to replace simila taxa names in the second data set. It relies on an algorithm that calculates the degree of similarity between names, usin the name from the first set which is most similar to that in the second data set.

When using the Synonymizer, keep in mind that order of selection matters. Always select the data set with the names you wish to use (the "real" name) *first*, and then, while holding down the CTRL key, click on the second data set with the taxa names you wish to change (the "synonym"). Then click on the **Synonymizer** button. A synonym data set wil be placed on the Data Tree panel under **Synonyms**. Each name in the data set selected second is now listed in the **TaxaSynonym** column. Next to this column is a **TaxaRealName** column listing the highest scoring match derived from the "real" name data set. The **MatchScore** column gives an indication of the amount of similarity between the two names (where 0 is no similarity and 1.0 is identity).




**Caution!** Before the synonyms are applied, we strongly encourage the user to check the match score, especially for those taxa with low match scores. To do that, the user selects the synonym file and clicks the "Synonymizer" button The incorrect matches, usually the ones with the lowest match scores, can be rejected at this point. Sorting on the match score column first makes this a fairly easy process.

In the event that some of the taxa are not interpreted correctly, matches can be modified manually. Select the taxa you wish to modify on the left side, and then choose a replacement taxa from the right side. Click the arrow button  to substitute the taxa. Taxa with no synonym can be identified by selecting then clicking "No Synonym". Click **OK** t save the changes.




Once it has been determined that the taxa names were matched correctly, the synonyms can be applied. With the synonyms selected, hold down the CTRL key while clicking on the second/synonym data set (the data set whose names you would like to change). Then once again click on the **Synonymizer** button to apply the new names to the data set.

## *3.6 Intersect Join*

### Command

```
./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt
-combine3 -input1 -input2 -intersect -export
group1_group2_intersect.hmp.txt -runfork1 -runfork2 -runfork3
```

This joins multiple data sets by the intersection of their taxa. Taxa must be present in both data sets to be included. Select multiple data sets using the CTRL key in conjunction with mouse clicks, and then click on the intersection butto to join the data sets. Because this function uses taxa names to join data sets, any variation in taxa names can prevent proper joining. Taxa names can be made uniform by using the “Synonymizer”.

## 3.7 Union Join

### Command

```
./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt
-combine3 -input1 -input2 -intersect -export group1_group2_union.hmp.txt
-runfork1 -runfork2 -runfork3
```

This joins multiple data sets by a union of their taxa. Missing data will be inserted if taxa are missing from one data se
Select multiple data sets using the CTRL key in conjunction with mouse clicks, and then click on the union button to joi
the data sets. Because this function uses taxa names to join data sets, any variation in taxa names can prevent proper
joining. Taxa names can be made uniform by using the "Synonymizer".

## 3.8 Merge Genotype Tables

### Command

```
./run_pipeline.pl -fork1 -h group1.hmp.txt -fork2 -h group2.hmp.txt
-combine3 -input1 -input2 -mergeAlignments -export
group1_group2_merge.hmp.txt -runfork1 -runfork2 -runfork3
```

This is the most complex merge function, and can be considered as a union join across both sites and taxa. (The actual -union join only works across taxa.) The resulting genotype table will contain all unique sites and all unique taxa from across the input datasets. If a specific site-taxon combination isn't present in any input dataset, the value is set to missing. If a specific site-taxon combination is present in more than one input file, the output will contain the last value processed. (That is, later values overwrite earlier values even if they conflict. There are plans to change this, but they have not been implemented yet.)

### Notes

- This maps to "Data -> Merge Genotype Tables" Menu on GUI.
- Error if duplicate site names in same file. (same as with other file loadings)
- Undefined taxa / site allele values are set to UNKNOWN.
- Duplicate taxa / site set to last Alignment processed.
- Sites are identified by Locus (chromosome), Physical Position, and Site Name

## 3.9 Separate

This separates the selected data set into it's components. For example, a genotype table would be separated into individual chromosomes.

### 3.10 Homozygous Genotype

This changes all heterozygous values to unknown (N).

# 4 Impute Menu

## 4.1 Genotypic Imputation

TASSEL5 contains two methods for imputing missing genotype information, one is a generalized approach suitable for all types of populations but optimized for those with higher inbreeding coefficients (FILLIN) and the other is specificall optimized for finding recombination break points in full-sib families (FSFHap). More information on these two methods can be found at:

Swarts et al. FSFHap (Full-Sib Family Haplotype Imputation) and FILLIN (Fast, Inbred Line Library ImputatioN) optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants, Plant Genome, *in review*.

**FSFHap (Full-Sib Family Haplotype Imputation):**

FSFHap imputes missing genotypes and corrects genotyping errors for inbred individuals in full-sib families. It is very useful for calling haplotypes in low-coverage GBS data. The individuals must be at least partially inbred because the method relies on finding inbred segments to identify haplotypes. It does not use the parent genotypes directly, but including the parents may be useful for interpreting the results. The algorithms used for imputation analyze one chromosome and family at a time. As a result, a pedigree file must be supplied that indicates which entries belong to which family. Also, input genotypes must contain data for only a single chromosome. If the genotype file contains multiple chromosomes, the chromosomes can be separated using the TASSEL -separate command.

*Pedigree File Format:*
The only file format specific to FSFHap is the pedigree file. The taxa names must exactly match names in the genotype data. If the genotype data contains taxa not included in the pedigree file, only individuals listed in the pedigree file will b analyzed. The input genotypes can be in any of the formats accepted by TASSEL. The pedigree file must contain the names of the individual taxa to be analyzed, the family to which each belongs, the parents, the parent contributions, and the average inbreeding coefficient. The first row in the file must be column headers. The values in the columns should b tab-delimited and are expected to be in the following order: family, taxon, parent1, parent2, parent1Contribution, parent2Contribution, F. The F value is not required but all other columns are.

Example:

| family | taxonName | parent1 | parent2 | contribution1 | contribution2 | F |
|---|---|---|---|---|---|---|
| fam1 | t0001 | par1 | par2 | 0.5 | 0.5 | .92 |
| fam1 | t0002 | par1 | par2 | 0.5 | 0.5 | .92 |
| ... | | | | | | |
| fam2 | t0201 | par1 | par3 | 0.5 | 0.5 | .92 |
| fam2 | t0202 | par1 | par3 | 0.5 | 0.5 | .92 |
| fam2 | t0203 | par1 | par3 | 0.5 | 0.5 | .92 |

The values for contribution1, contribution2, and F are family means. Those values are read from the first line for a famil
only and then applied to the entire family.

*Using the command line for FSFHap:*
FSFHap consists of three TASSEL plugins, CallParentAllelesPlugin, ViterbiAlgorithmPlugin, and WritePopulationAlignmentPlugin, which are called sequentially. A typical command for running FSFHap is as follows (replace items in <> with actual parameter values) for a genotype containing a single chromosome:

```
run_pipeline.pl -h <genotypeFilename> -CallParentAllelesPlugin -p <pedigreeFilename> -m 0.9
-r 0.5 -logfile <logFilename> -endPlugin -ViterbiAlgorithmPlugin -g true -endPlugin
-WritePopulationAlignmentPlugin -f <outputFilename> -m false -o parents -endPlugin
```

For a genotype file containing multiple chromosomes:

```
run_pipeline.pl -h <genotypeFilename> -separate -CallParentAllelesPlugin -p
<pedigreeFilename> -m 0.9 -r 0.5 -logfile <logFilename> -endPlugin -ViterbiAlgorithmPlugin
-g true -endPlugin -WritePopulationAlignmentPlugin -f <outputFilename> -m false -o parents
-endPlugin
```

*Options for CallParentAllelesPlugin:*
Options taking a parameter value specified by Value = []:

| | |
|---|---|
| -p or -pedigrees | the pedigree file. Value = [filename] |
| -w or -windowSize | the number of SNPs to examine for each LD cluster. Value = [integer] (default = 50) |
| -r or -minR | minimum R used to filter SNPs on LD Value = [number between 0 and 1]. (default = 0.2, use 0 for no ld filter) |
| -m or -maxMissing | maximum proportion of missing data allowed for a SNP Value = [number between 0 and 1]. (default = 0.9) |
| -f or -minMaf | minimum minor allele frequency used to filter SNPs. If negative, filters on expected segregation ratio from parental contribution. Value = [number between 1 and -1]. (default = -1) |
| -b or -bc1 | use BC1 specific filter. Value = [true or false] (default = true) |
| -n or -bcn | use multiple backcross specific filter. Value = [true or false] (default = false) |
| -logfile | the name of a file to which all logged messages will be printed. Value = [filename]. |

Options not taking a parameter value:

| | |
|---|---|
| -cluster | use the cluster algorithm. minMaf defaults to 0.05. |
| -subpops | filter sites for heterozygosity in subpopulations. |
| -nohets | delete het calls from original data before imputing. |
| -windowld | use the window ld algorithm for finding parent haplotypes |

The “-cluster”, “-subpops”, “-nohets”, and “-windowld” options do not take parameters but only act as flags that include certain features in the analysis. Of those, cluster and windowld are the most useful. When the -cluster option is used, a different algorithm is used that does a better job of handling residual heterozygosity in the parents. However, it does not perform well for partially inbred RILs that have only been self-pollinated for one or two generations. If the RILs being imputed are F2’s or F3’s, the “-cluster” option should not be used. The “-subpops” option should only be used when imputing families of the NAM population developed by the Maize Diversity Project. The “-nohets” option

was included to test whether or not erroneous het calls result in too many hets being imputed. It appears to have only a small effect on the outcome. The -windowld algorithm handles F2 and later populations effectively, but can have problems when parents have some residual heterozygosity.

It is recommended that the -logfile option be used. The output can be used to identify and diagnose possible problems. The "-bcn true" should be used for populations with two or more backcrosses. However, using the "-bc1" option is not necessary as the default behavior is usually best.

*Options for ViterbiAlgorithmPlugin:*

| | |
|---|---|
| -g or -fillgaps | if true then missing values flanked by SNPs from the same parent will be imputed to that parent, false otherwise. Value = [true or false] (default = true) |
| -h or -phet | expected frequency of heterozygous loci. Used only if the inbreeding coefficient is not specified in the pedigree file. Value = [number between 0 and 1] (default = 0.07) |

*Options for WritePopulationAlignmentsPlugin:*

Required:

| | |
|---|---|
| -f or -file | The base file name for the ouput. .hmp.txt will be appended. Value = [filename] |

Optional:

| | |
|---|---|
| -m or -merge | if true then families are merged into a single file, if false then each family is output to a separate file. Value = [true or false] (default = false) |
| -o or -outputType | if value = parents then output parent calls, if value = nucleotides then output nucleotides, if value = both then output both in separate files (default = both) |
| -d or -diploid | if true output is AA/CC/AC, if false output is A/C/M. Value = [true or false] (default = false) |
| -c or -minCoverage | the minimum coverage for a monomorphic snp to be included in the nucleotide output. Value = [number between 0 and 1] (default = 0.1) |
| -x or -maxMono | the maximum minor allele frequency used to call monomorphic snps (default = 0.01) |

For individual families, only polymorphic SNPs are imputed. When merge = false, only those SNPs appear in the output. When merge = true, SNPs that are polymorphic in any family will be written to output. For any site, if SNP coverage is high enough in a family to determine with confidence that it is monomorphic for that family, then all individuals in that family will be imputed to the monomorphic value at that site. The -minCoverage and -maxMono options are used to determine thresholds for determining whether a site will be called monomorphic in a family. If either of the options is set to a value of NaN, then missing values at monomorphic sites will not be imputed.

**FILLIN (Fast, Inbred Line Library ImputatioN): The generalized approach**

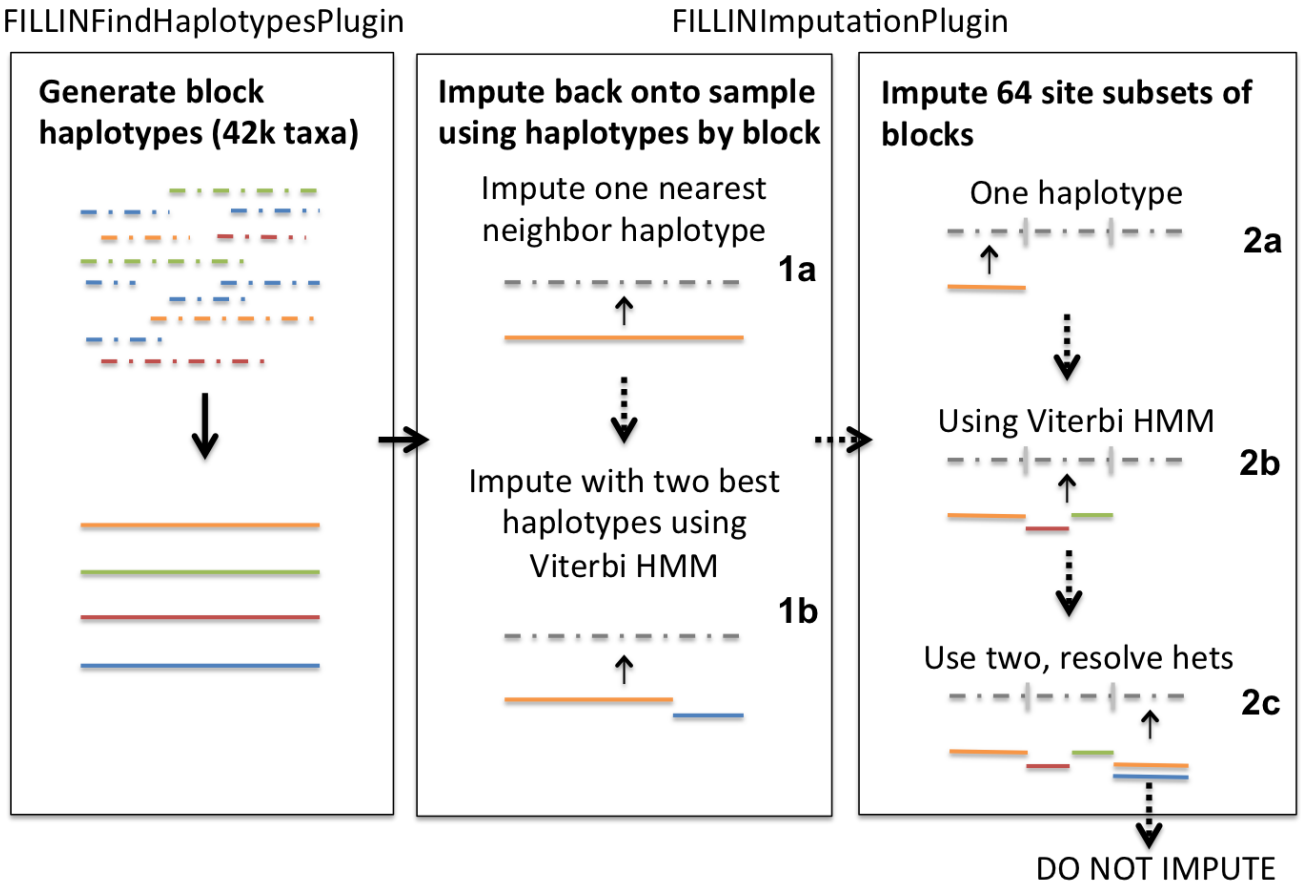
FILLIN imputes missing genotypes in two steps, 1) haplotype generation (FILLINFindHaplotypesPlugin) and 2) imputation of the resulting haplotypes back onto the target samples (FILLINImputationPlugin).

Haplotypes are generated by collapsing low coverage but inbred segments that share identity by state to an optionally user-supplied threshold value by site window (default: 8k); this is performed by the first plugin, FILLINFindHaplotypesPlugin. Because short IBD segments may be replicated widely within a species, even between diverse individuals, we recommend supplying all the information available within a species for this step.

The second plugin, FILLINImputationPlugin, uses these haplotypes to impute missing genotypes in target individuals. I does so in multiple steps, first looking for haplotypes that match the minor alleles to a threshold within the whole site window (1a in schematic below) and, if this fails, looks for two haplotypes to explain the site window and, assuming thi

represents a recombination break point between two inbred haplotypes, uses a Viterbi HMM algorithm to model the recombination breakpoints (2a). If two haplotypes cannot be found to explain the whole site window, the algorithm next searches for haplotypes to explain a smaller focus window within the site window centered on 64 sites at a time and searching to the right and left until enough informative minor alleles are found. It does this by first looking for one haplotype to a threshold (2a), then two modeling a recombination break between inbred segments (2b), then finally, to a higher threshold, looks for two haplotypes and models the 64 focus site window as heterozygous, combining the two haplotypes together. The thresholds for 2a-c are also set differently based on whether the whole sequence of the target taxon is above or below a user supplied heterozygosity threshold. For taxon considered outbred (above the threshold), 2b the Viterbi option is never used because it is more likely in an outbred taxon that if two haplotypes explain a segment it is heterozygous for those two haplotypes. If the algorithm cannot find haplotypes to satisfy any of these threshold requirements, the segment will not be imputed. The thresholds for the focus block imputation are set based on the mxInbErr and mxHybErr values entered (or defaults):

| | Below mxHet (inbred) | Above mxHet (outbred) |
|---|---|---|
| 2a | 3/10*mxInbErr | 1/10*mxInbErr |
| 2b | ⅓*mxHybErr | 0 |
| 2c | mxInbErr | mxInbErr |




*Running FILLIN:*

FILLIN consists of two TASSEL plugins, FILLINFindHaplotypesPlugin and FILLINImputationPlugin, which are called sequentially. If you would like to mask your data and calculate accuracy, use the -accuracy flag for

FILLINImputationPlugin. If imputing maize, a donor file of haplotypes from 40k+ taxa can be found on the Panzea website (http://www.panzea.org/lit/data_sets.html). FILLIN can be run either within the TASSEL GUI or through the command line. The options are the same for both.

A typical command sequence for running FILLIN through the command line is as follows (replace items in <> with actual parameter values):

```
run_pipeline.pl -FILLINFindHaplotypesPlugin -hmp <genotypeFilename> -o <outDonorDir>
run_pipeline.pl -FILLINImputationPlugin -hmp <genotypeFilename> -d <donorDir> -o
<outFile.hmp.txt.gz>
```

To run FILLIN from the GUI go to Impute->FILLINFindHaplotypesPlugin or FILLINImputationPlugin

*Options for FILLINFindHaplotypesPlugin:*

-hmp <Target file> :
Input genotypes to generate haplotypes from. Usually best to use all available samples from a species. Accepts all file types supported by TASSEL5. (required)

-o <Donor dir/file basename> :
Output file directory name, or new directory path; Directory will be created, if doesn't exist. Outfiles will be placed in the directory and given the same name and appended with the substring '.gc#s#.hmp.txt' to denote chromosome and section (required)

-mxDiv <Max divergence from founder> :
Maximum genetic divergence from founder haplotype to cluster sequences (Default: 0.01)

-mxHet <Max heterozygosity of output haplotypes> :
Maximum heterozygosity of output haplotype. Heterozygosity results from clustering sequences that either have residual heterozygosity or clustering sequences that do not share all minor alleles. (Default: 0.01)

-minSites <Min sites to cluster> :
The minimum number of sites present in two taxa to compare genetic distance to evaluate similarity for clusterin (Default: 50)

-mxErr <Max combined error to impute two donors> :
The maximum genetic divergence allowable to cluster taxa (Default: 0.05)

-hapSize <Preferred haplotype size> :
Preferred haplotype block size in sites (minimum 64); will use the closest multiple of 64 at or below the supplied value (Default: 8192)

-minPres <Min sites to test match> :
Minimum number of present sites within input sequence to do the search (Default: 500)

-maxHap <Max haplotypes per segment> :
Maximum number of haplotypes per segment (Default: 3000)

-minTaxa <Min taxa to generate a haplotype> :
Minimum number of taxa to generate a haplotype (Default: 2)

-maxOutMiss <Max frequency missing per haplotype> :
Maximum frequency of missing data in the output haplotype (Default: 0.4)

-nV <true | false> :
Supress system out (Default: false)

-extOut <true | false> :
Details of taxa included in each haplotype to system out (Default: false)

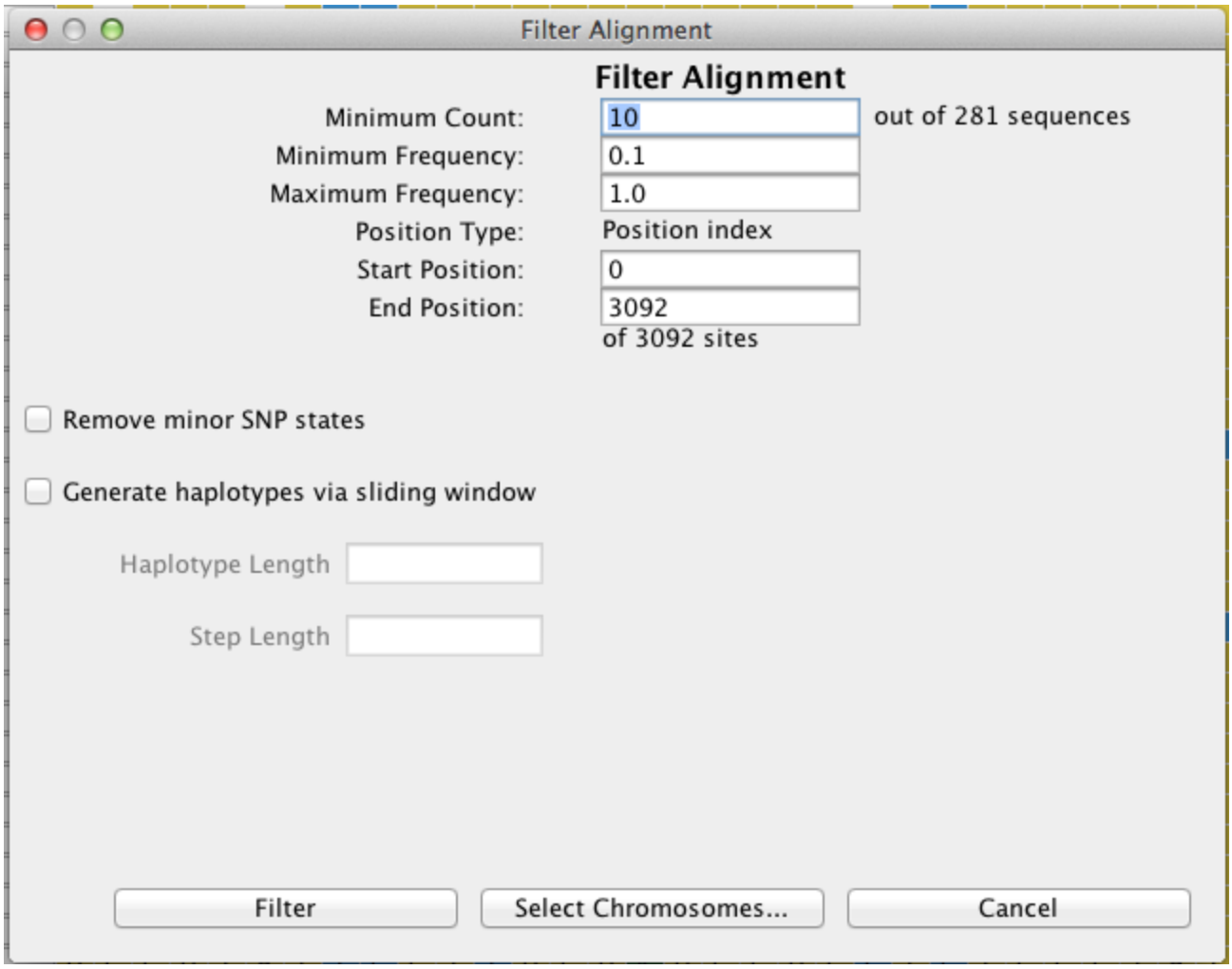
*Options for FILLINImputationPlugin:*
-hmp <Target file> :
Input HapMap file of target genotypes to impute. Accepts all file types supported by TASSEL5 (required)
-d <Donor Dir> :
Directory containing donor haplotype files from output of FILLINFindHaplotypesPlugin. All files with '.gc' in th
filename will be read in, only those with matching sites are used (required)
-o <Output filename> :
Output file; hmp.txt.gz and .hmp.h5 accepted. (required)
-hapSize <Preferred haplotype size> :
Preferred haplotype block size in sites (use same as in FILLINFindHaplotypesPlugin) (Default: 8000)
-hetThresh <Heterozygosity threshold> :
Threshold per taxon heterozygosity for treating taxon as heterozygous (no Viterbi, het thresholds). (Default
0.01)
-mxInbErr <Max error to impute one donor> :
Maximum error rate for applying one haplotype to entire site window (Default: 0.01)
-mxHybErr <Max combined error to impute two donors> :
Maximum error rate for applying Viterbi with to haplotypes to entire site window (Default: 0.003)
-mnTestSite <Min sites to test match> :
Minimum number of sites to test for IBS between haplotype and target in focus block (Default: 20)
-minMnCnt <Min num of minor alleles to compare> :
Minimum number of informative minor alleles in the search window (or 10X major) (Default: 20)
-mxDonH <Max donor hypotheses> :
Maximum number of donor hypotheses to be explored (Default: 20)
-hybNN <true | false> :
If true, uses combination mode in focus block, else does not impute (Default: true)
-ProjA <true | false> :
Create a projection alignment for high density markers (Default: false)
-impDonor <true | false> :
Impute the donor file itself (Default: false)
-nV <true | false> :
Supress system out (Default: false)
*Options for calculating accuracy*
-accuracy <true | false> :
Masks input file before imputation and calculates accuracy based on masked genotypes (Default: false)
-propSitesMask <Proportion of genotypes to mask if no depth> :
Proportion of genotypes to mask for accuracy calculation if depth not available (Default: 0.01)
-depthMask <Depth of genotypes to mask> :
Depth of genotypes to mask for accuracy calculation if depth information available (Default: 9)
-propDepthSitesMask <Proportion of depth genotypes to mask> :
Proportion of genotypes of given depth to mask for accuracy calculation if depth available (Default: 0.2)

# 5 Filter Menu

## *5.1 Sites*

The genotype table can be filtered in several ways. For example, monomorphic sites can be eliminated, and regions of a

sequence can be eliminated.




**Minimum Count** - the minimum number of taxa in which the site must have been scored to be included in the filtere data set (GAP or missing data do not count).
**Minimum Frequency** - the minimum frequency of the minority polymorphisms for the site to be included in the filtere data set.
**Start Position, End Position** – establishes the range of sites for filtering.
**Extract Indels** - if selected, indels are extracted from the alignment. If not selected, only point substitutions are extracted.
**Remove minor SNP states** – converts tertiary and rarer states to missing data ("?"), thereby forcing sites to have only two types of segregating sites at a locus. This may help remove sequencing errors.
**Generate haplotypes via sliding window** – creates haplotypes from an ordered set of SNPs.

**Example Pipeline Command that removes SNPs with MAF (Minimum Allele Frequency) less than 5%**

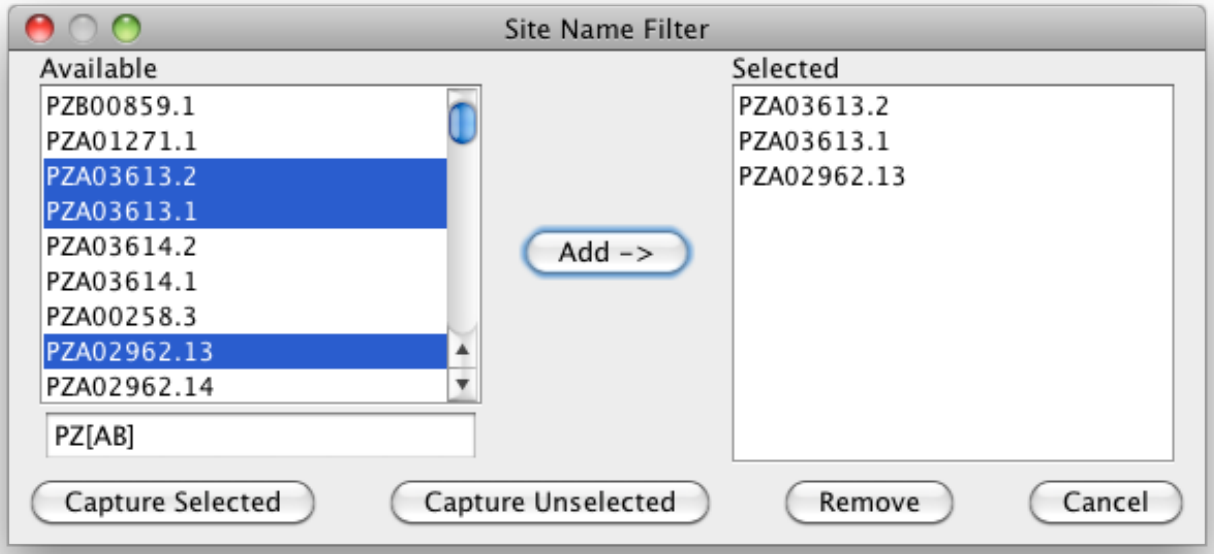
```
run_pipeline.pl -fork1 -h mdp_genotype.hmp.txt -filterAlign
-filterAlignMinFreq 0.05 -export filtered_genotype -runfork1
```

## 5.2 Site Names

First select the genotypic data from the data tree. The resulting dialog displays the site names associated with the selected data. By using either the CTRL or SHIFT key in conjunction with the mouse, the user can select or deselect site names. Once desired site names have been moved to the "Selected" window using the "Add ->" button, the "Capture Selected" or "Capture Unselected" buttons will create a new data set containing only the desired site names.

Using the search box…

- * is the wildcard.
- * is always implied at end of search string.
- Search string is case sensitive. For example: use [Aa]bc to match site names beginning with Abc or abc.
- PZ[AB] Will match anything starting with PZA or PZB.



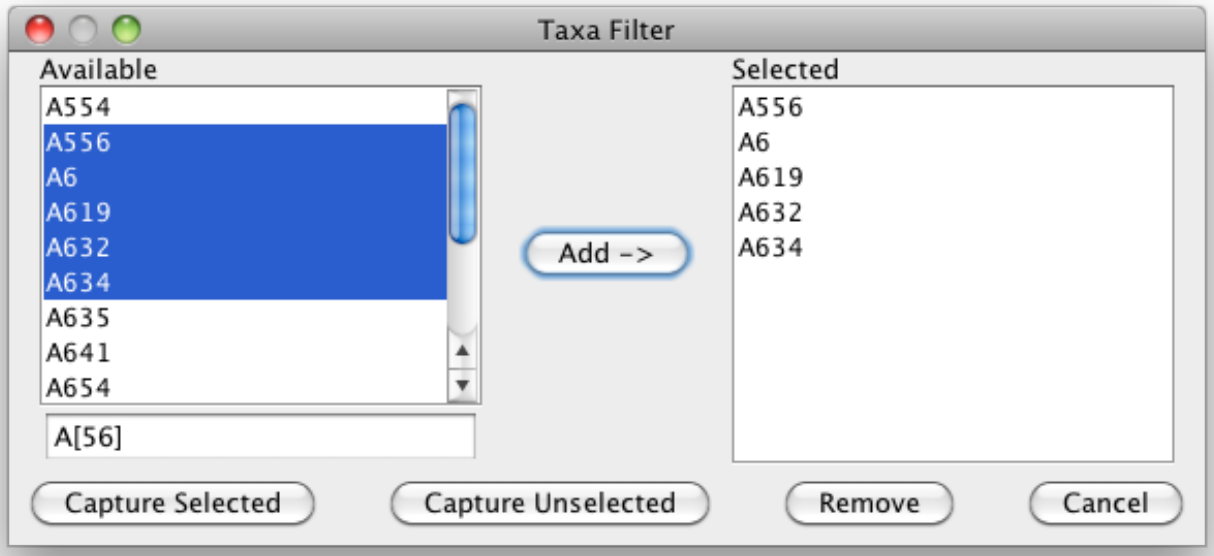

## 5.3 Taxa Names

First select the genotypic, phenotypic, or population structure data from the data tree. The resulting dialog displays the taxa associated with the selected data. By using either the CTRL or SHIFT key in conjunction with the mouse, the user can select or deselect taxa. Once desired taxa have been moved to the "Selected" window using the "Add ->" button, the "Capture Selected" or "Capture Unselected" buttons will create a new data set containing only the desired taxa.

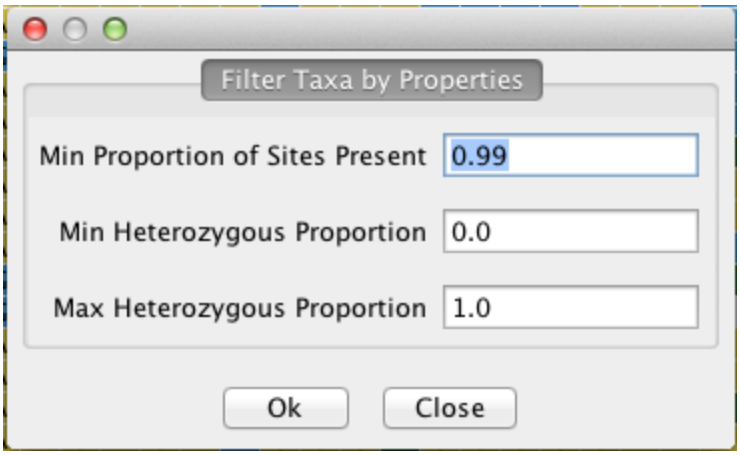
Using the search box…

- * is the wildcard.
- * is always implied at end of search string.
- Search string is case sensitive. For example: use [Aa]bc to match taxa beginning with Abc or abc.

- A[56] Will match anything starting with A5 or A6




## 5.4 Taxa



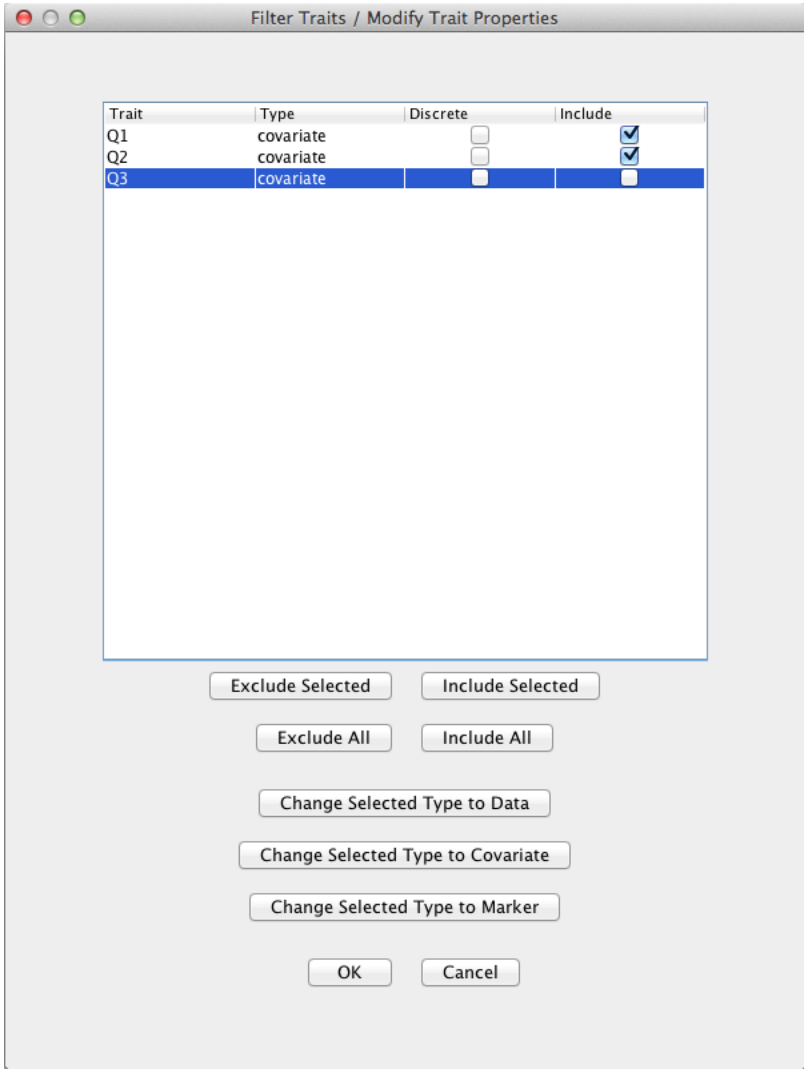

## 5.5 Traits

Clicking the “Traits” button on the “Data” toolbar launches the Trait Filter dialog. This dialog is used with numerical da
sets to (1) change the trait type, (2) view, but not change whether the trait is discrete or continuous and (3) drop one or
more traits from the data set. In addition, the dialog can be used to view the trait properties without changing them. I
the “OK” button is clicked, a new data set is created that incorporates the changes, the original data set remains
unchanged, and the dialog closes. If the “Cancel” button is clicked no data set is created, the original data set remains
unchanged, and the dialog closes.
Allowable trait types are data, covariate, factor and marker. Generally, data and covariate traits will be continuous (not
discrete) and factor will be discrete. Markers in a numerical data set will be continuous. Discrete valued markers are

better imported as genotypes and filtered using the “Sites” filter.

Clicking “Exclude All” unchecks the “Include” box for all traits. Clicking “Include All” checks the “Include” box for traits. The “Exclude Selected” and “Include Selected” buttons do the same thing for traits that have been highlighted b selecting them with the mouse. Type can be changed for individual traits by selecting a value in the drop down box in th type column for that trait. Type can be changed for multiple traits by selecting those traits then clicking one of th “Change Selected Type to …” buttons.

**Important:** Once a numerical data set has been joined with genotypes, it can no longer be modified using the trait filte function.



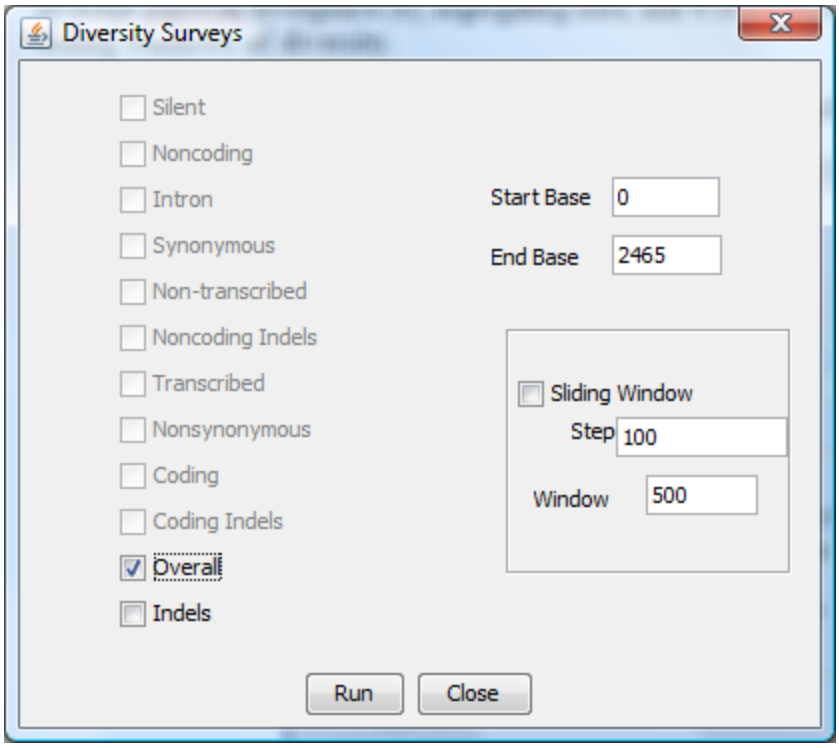

# 6 Analysis Menu

## *6.1 Diversity*

This executes a basic diversity analysis.

Average pairwise divergence (π), segregating sites, and θ estimates (4Nμ) can be calculated, as well as sliding windows of diversity.

To run a diversity analysis, click on a raw sequence alignment, and then select **Analysis** -> **Diversity**.



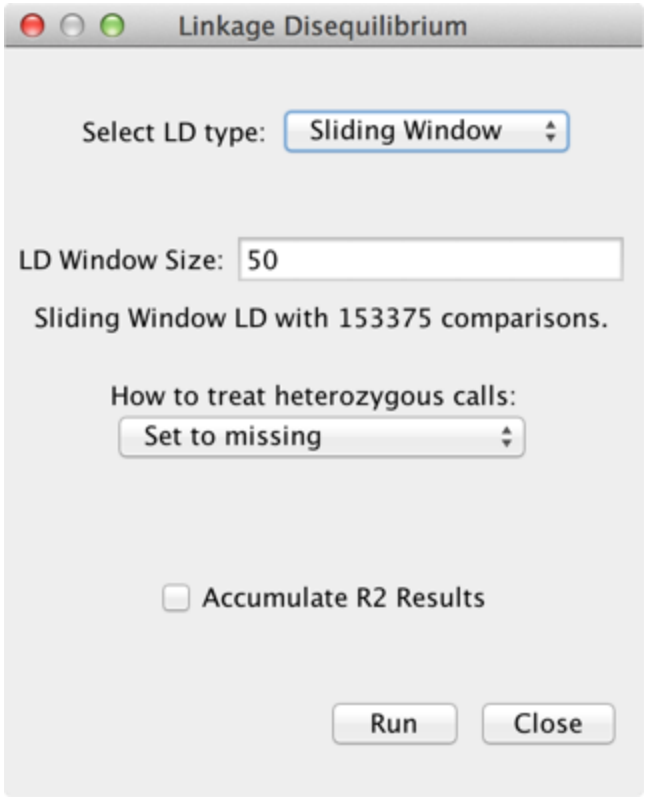

In the resulting Diversity Surveys dialog box, the various site classes available for analysis are listed on the left. If t sequence has no annotation, then only the "Overall" and "Indels" options will be active.
A sliding window of diversity can also be calculated across the region. To produce a sliding window, check the box next to "Sliding Window," and then enter the desired step size and size of the sliding window.

Results can be plotted using **Results** -> **Chart** or viewed in a table via **Results -> Table**.

## *6.2 Linkage Disequilibrium*

This generates a linkage disequilibrium data set from SNP data.

NOTE: It is important to use only filtered data sets (apply **Filter -> Sites** first) when estimating linkage disequilibriun as a raw alignment with numerous invariant bases will take a very long time and consume a large amount of memory t calculate.




Linkage disequilibrium between any set of polymorphisms can be estimated by clicking on a filtered set o polymorphisms and then using **Analysis □ Link. Diseq.** At this time, *D'*, *r2* and *P*-values will be estimated. The current version calculates LD between haplotypes with known phase only (unphased diploid genotypes are not supported; see PowerMarker or Arlequin for genotype support).

> **D'** is the standardized disequilibrium coefficient, a useful statistic for determining whether recombination o homoplasy has occurred between a pair of alleles.
>
> $r^2$ represents the correlation between alleles at two loci, which is informative for evaluating the resolution o association approaches.

*D'* and *r2* can be calculated[21] when only two alleles are present. If multiple alleles are present, a weighted average of *D* or *r2* is calculated between the two loci[22]. This weighted average is determined by calculating *D'* or *r2* for all possible combinations of alleles, and then weighting them according to the allele's frequency. *Note: It is not entirely certain that this procedure fully accounts for allele number effects.*

**P-values** are determined by two methods. If only two alleles are present at both loci, then a two-sided Fisher's Exact test is calculated. *Note: Previous editions of TASSEL used a one-sided test, but TASSEL version 1.0.8 and later use a two-sided test.*

If more than two alleles are present, permutations are used to calculate the proportion of permuted gamete distributions that are less probable then the observed gamete distribution under the null hypothesis of independence[21] .

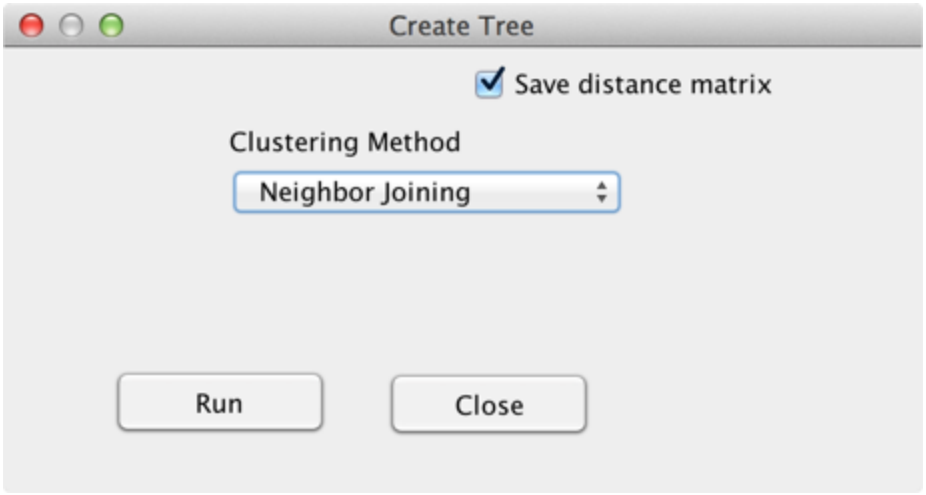
When calculating linkage disequilibrium, users have the option of employing "**Rapid Permutations."** If this option selected, the algorithm will compute either a fixed number of permutations or run until 10 permutations are found that a

more significant than the observed P-value. While this slightly reduces P-values, it also saves a large amount o computational time. If an unbiased p-value is desired, then the user must unselect the **"Rapid Permutations"** check box.

"Full Matrix LD" calculates LD for every combination of sites in the alignment. "Sliding Window LD" calculates LD sites within a window of sites surrounding the current site. The LD Window Size determines the width of the windo on one side of the current site.

Linkage disequilibrium results can be plotted using **Results** -> **LD Plot** or viewed in a table via (**Results -> Table**).

## 6.3 Cladogram




This function generates a tree or cladogram data set.

TASSEL produces neighbor-joining trees using only simple parsimony substitution models.

To retrieve cladogram data, first select genotypic data from the Data Tree and then click **Analysis -> Cladogram**. The resulting tree data and the corresponding matrix will appear as separate data sets on the Data Tree.

Results can be plotted using **Results** -> **Archaeopteryx Tree.**

## 6.4 Kinship

This function generates a kinship matrix from a genotype. To do so, first highlight SNP data then click on the "**Analysis/Kinship**" submenu. The resulting dialog box will then provide the option to select "scaled IBS" or "pairwise IBS". Clicking "OK" generates a kinship matrix.

When a genotype file is selected and "pairwise IBS", each element i,j of the kinship matrix that is generated is equal t the proportion of the SNPs which are different between taxon i and taxon j. Distance is calculated for each pair of taxa, ignoring any sites that have a missing value for one of the taxa. The distance matrix is converted to a similarity matrix

subtracting all values from 2 then scaling so that the minimum value in the matrix is 0 and the maximum value is Kinship can be derived from a set of random SNP data (a minimum of several hundred SNPs spread over the whole genome is recommended). This ad-hoc rescaling method was implemented in an earlier version of TASSEL in order to provide a reasonable estimate of additive genetic variance, but tends to overestimate that value. Rescaling does not affect its use for correcting for population structure. It only affects the estimate of additive genetic variance and consequently, heritability.

To provide a better estimate of addivitive genetic variance, an alternative method can be used by selecting "scaled IBS" This method (from Endelman and Jannink, 2012) codes genotypes as 2, 1, or 0, equal to the count of one of the alleles at that locus. It then replaces missing genotype values with the average genotypic score at that locus before estimating a relationship matrix. Other methods of imputing genotypes prior to calculating Kinship may provide a better result. Fo instance, rather than using this default treatment of missing values, using the numerical genotype method followed imputation described in section 3.3 before running Kinship is a reasonable alternative. When using numerical genotype Kinship always applies the "scaled IBS" method.

Users may also load their own kinship data using **Data □ Load**. Kinship matrices can be calculated using the SPAGeDi software package (http://www.ulb.ac.be/sciences/ecoevol/spagedi.html). Comparisons of methods for calculating kinship can be found in the literature (*e.g.* Stich et al. 2008).

## *6.5 GLM (General Linear Model)*

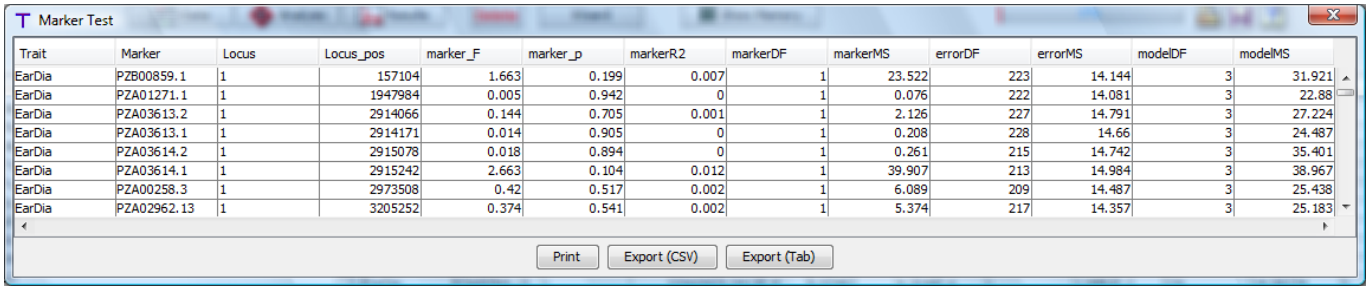
This function performs association analysis using a least squares fixed effects linear model.

TASSEL utilizes a fixed effects linear model to test for association between segregating sites and phenotypes. The analysis optionally accounts for population structure using covariates that indicate degree of membership in underlyi populations. A main effects only model is automatically built using all variables in the input data. A separate model built and solved for each trait and marker combination. Any factors, covariates, reps or locations are included in every model as main effects. How the data is used must be defined either in the input data files or using the **Trait Filter** afte the data has been imported but before it has been joined with a genotype.

General Linear Model (GLM) can be run using a numeric data set only or using numeric data joined to genotype data. If only numeric data is selected, best linear unbiased estimates (BLUEs or least square means) will be generated for the taxa for each trait. [Note: only factors and covariates intended to control field variation should be included at this stage Population structure covariates which are intended to control for marker effects should only be included when markers are also in the analysis.] If numeric data with genotypes are analyzed, each trait by marker combination will be tested and two reports will be produced, one containing trait by marker F-tests and the other containing allele estimates.

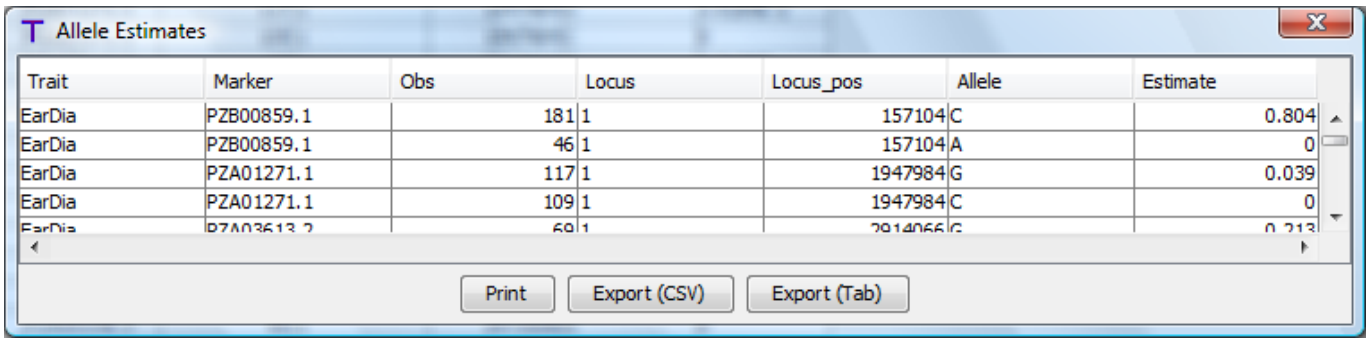
To run GLM, select a data set and then click the GLM button. A dialog box will pop-up to allow the user to indicate that a permutation test should be run and to allow the number of permutations to be changed. The permutation test will be run using the method suggested by Anderson and Ter Braak (2003), which calculates the predicted and residual values of the reduced model (contained all terms except markers) then permutes the residuals and adds them to the predicted values. When the GLM options dialog is closed, the user is presented with a dialog allowing the output to be saved to a file rather than stored in memory and displayed by TASSEL. This option is useful when the output is expected to be very large and risks exceeding available RAM.

The following table shows an example of the Marker Test output as viewed with Results/Table:




| Trait | Marker | Locus | Locus_pos | marker_F | marker_p | markerR2 | markerDF | markerMS | errorDF | errorMS | modelDF | modelMS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EarDia | PZB00859.1 | 1 | 157104 | 1.663 | 0.199 | 0.007 | 1 | 23.522 | 223 | 14.144 | 3 | 31.921 |
| EarDia | PZA01271.1 | 1 | 1947984 | 0.005 | 0.942 | 0 | 1 | 0.076 | 222 | 14.081 | 3 | 22.88 |
| EarDia | PZA03613.2 | 1 | 2914066 | 0.144 | 0.705 | 0.001 | 1 | 2.126 | 227 | 14.791 | 3 | 27.224 |
| EarDia | PZA03613.1 | 1 | 2914171 | 0.014 | 0.905 | 0 | 1 | 0.208 | 228 | 14.66 | 3 | 24.487 |
| EarDia | PZA03614.2 | 1 | 2915078 | 0.018 | 0.894 | 0 | 1 | 0.261 | 215 | 14.742 | 3 | 35.401 |
| EarDia | PZA03614.1 | 1 | 2915242 | 2.663 | 0.104 | 0.012 | 1 | 39.907 | 213 | 14.984 | 3 | 38.967 |
| EarDia | PZA00258.3 | 1 | 2973508 | 0.42 | 0.517 | 0.002 | 1 | 6.089 | 209 | 14.487 | 3 | 25.438 |
| EarDia | PZA02962.13 | 1 | 3205252 | 0.374 | 0.541 | 0.002 | 1 | 5.374 | 217 | 14.357 | 3 | 25.183 |

In addition to displaying the F-statistics and p-values for the requested F-tests, the table also contains markerR2, mean squares (MS) and degrees of freedom (DF) for the marker effect, for the model (corrected for the mean), and for error. If taxa are replicated (across reps or environments), then the markers are tested using the taxa within marker mean square. If taxa are unreplicated, then the residual mean square is used. MarkerR2 is the marginal R-squared for the marker calculated as SS Marker (after fitting all other model terms) / SS Total, where SS stands for sum of squares. The following table shows an example of the Allele Estimates output as viewed with Results/Table:




| Trait | Marker | Obs | Locus | Locus_pos | Allele | Estimate |
|---|---|---|---|---|---|---|
| EarDia | PZB00859.1 | 181 | 1 | 157104 | C | 0.804 |
| EarDia | PZB00859.1 | 46 | 1 | 157104 | A | 0 |
| EarDia | PZA01271.1 | 117 | 1 | 1947984 | G | 0.039 |
| EarDia | PZA01271.1 | 109 | 1 | 1947984 | C | 0 |
| EarDia | PZA03613.2 | 69 | 1 | 2914066 | G | 0.213 |

For each marker and trait combination, each marker allele is listed along with the number of observations for taxa carrying that allele (Obs), the locus (usually chromosome) and locus position of that marker, the allele, and the estimat of the effect of that allele. Because of the way that GLM codes alleles, the last allele estimate for a marker is always zero and the other allele estimates are relative to that.

## 6.6 *MLM (Mixed Linear Model)*

This conducts association analysis via a mixed linear model (MLM).

A mixed model is one which includes both fixed and random effects. Including random effects gives MLM the ability incorporate information about relationships among individuals. When a genetic marker based kinship matrix (K) is use jointly with population structure (Q), the "Q+K" approach improves statistical power compared to "Q" only[9]. MLM can be described in Henderson's matrix notation[23] as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathrm{e}$$

where **y** is the vector of observations; **β** is an unknown vector containing fixed effects, including genetic marker an population structure (Q); **u** is an unknown vector of random additive genetic effects from multiple background QTL fo individuals/lines; **X** and **Z** are the known design matrices; and **e** is the unobserved vector of random residual. The **u** an **e** vectors are assumed to be normally distributed with null mean and variance of

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

where $\mathbf{G} = \sigma_a^2\mathbf{K}$ with $\sigma_a^2$ as the additive genetic variance and **K** as the kinship matrix. Homogeneous variance is assumed for the residual effect which means $\mathbf{R}=\mathbf{I}\sigma_e^2$, where $\sigma_e^2$ is the residual variance. The proportion of genetic variance over the total variance is defined as heritability ($h^2$).

When K is derived from pedigrees, the elements of K equal 2*Probability(IBD), where IBD means that two alleles drawn at random are identical by descent. Generally, K calculated from markers is an IBS matrix. The resulting multiplier is then not $\sigma_a^2$ but some unknown constant times $\sigma_a^2$. Some methods for calculating K, such as those implemented in SPaGEDI, actually use markers to develop an estimate of the IBD relationship matrix. For those values of K, the resulting variance estimate can be considered an estimate of $\sigma_a^2$ as long as the assumptions of the method used to derive K are not violated for the population being analyzed. One implication is that two different K matrices may give very different estimates of $\sigma_a$ and heritability yet produce the same model fit and test of marker association.

TASSEL implements several methods to improve statistical power and reduce computing time. The Restricted Maximum Likelihood (REML) estimates of $\sigma_a^2$ and $\sigma_e^2$ are obtained through the Efficient Mixed-Model Association (EMMA) algorithm[24] which is much faster than the expectation and maximization (EM) algorithm[25].

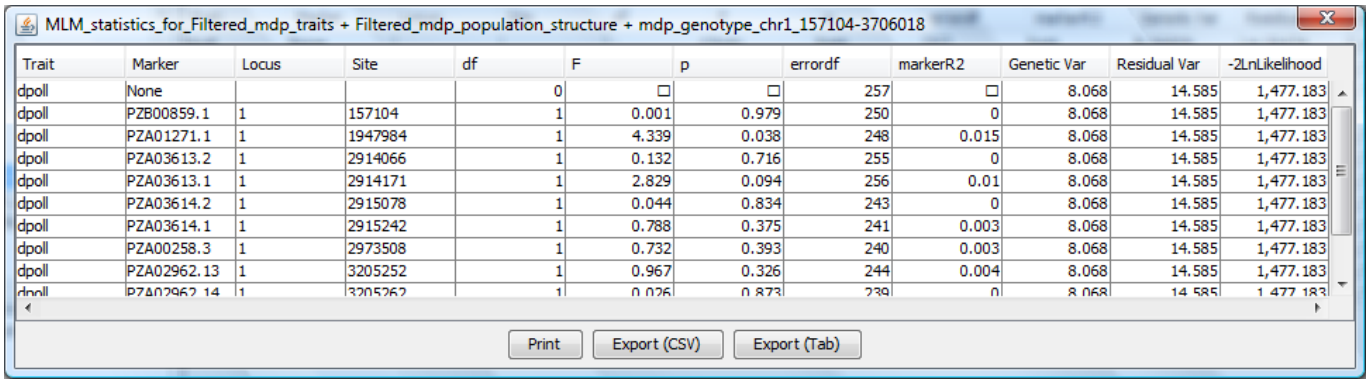
TASSEL also implements a method called compression which reduces the dimensionality of the kinship matrix to reduce computational time and improve model fitting. When MLM is used without compression (compression = 1), each taxon belongs to its own group. At the other extreme, GLM can be interpreted as maximum compression (compression = n) with all taxa in a single group. In that case, it is not possible to estimate the random effect independently of error and $\sigma_a^2$ is absorbed into $\sigma_e^2$. Between these two extremes, taxa can be grouped using cluster analysis based on kinship. When n individuals are compressed into s clusters (groups), the kinship among individuals is replaced with the kinship among groups. At some grouping levels, dependent on the trait and population being analyzed, this compressed MLM has improved statistical power compared to the regular MLM[4]. The optimum grouping with the best model fit for MLM without fitting genetic markers has the best statistical power for an association test of markers[4]. TASSEL allows users to specify the compression level (average number of individuals per group), or to have the program determine the optimum grouping.

Similar to GLM, MLM performs an association test for each combination of traits and markers. TASSEL provides users several options: 1) to estimate genetic and residual variance for each combination; 2) to get these estimates once for each trait without fitting genetic markers and then to use those estimates to test markers; 3) to use a prior heritability estimate provided by the user. The second option, named P3D (population parameters previously determined), has the same statistical power as the first option[4]. Using the P3D method or using a prior heritability can be much faster than calculating heritability for each marker.

Using MLM is very similar to using GLM. The difference is that in addition to choosing the joint data set (or numerical data set), kinship data must also be highlighted before clicking the MLM button to show the MLM option dialog. The option of “No Compression” is the regular MLM which is equivalent to “Custom level=1”. For data sets with large

numbers of taxa, the optimal compression option may be considerably slower than no compression or user supplied compression. This is because the algorithm solves the model once for each of a series of compression levels in order to determine the optimal one.

All MLM analyses create two output tables, model statistics and model effects. If compression is used, the analysis creates three tables.

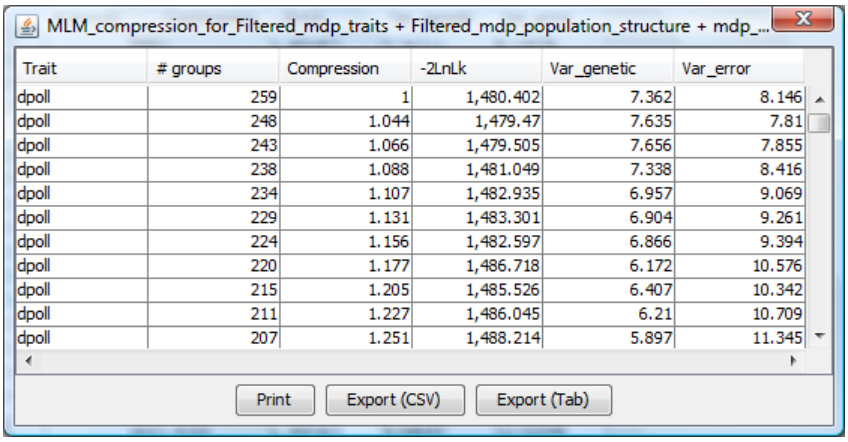

MLM_statistics_for_Filtered_mdp_traits + Filtered_mdp_population_structure + mdp_genotype_chr1_157104-3706018

| Trait | Marker | Locus | Site | df | F | p | errordf | markerR2 | Genetic Var | Residual Var | -2LnLikelihood |
|---|---|---|---|---|---|---|---|---|---|---|---|
| dpoll | None | | | 0 | | | 257 | | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZB00859.1 | 1 | 157104 | 1 | 0.001 | 0.979 | 250 | 0 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA01271.1 | 1 | 1947984 | 1 | 4.339 | 0.038 | 248 | 0.015 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA03613.2 | 1 | 2914066 | 1 | 0.132 | 0.716 | 255 | 0 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA03613.1 | 1 | 2914171 | 1 | 2.829 | 0.094 | 256 | 0.01 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA03614.2 | 1 | 2915078 | 1 | 0.044 | 0.834 | 243 | 0 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA03614.1 | 1 | 2915242 | 1 | 0.788 | 0.375 | 241 | 0.003 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA00258.3 | 1 | 2973508 | 1 | 0.732 | 0.393 | 240 | 0.003 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA02962.13 | 1 | 3205252 | 1 | 0.967 | 0.326 | 244 | 0.004 | 8.068 | 14.585 | 1,477.183 |
| dpoll | PZA02962.14 | 1 | 3205262 | 1 | 0.026 | 0.873 | 239 | 0 | 8.068 | 14.585 | 1,477.183 |

Print Export (CSV) Export (Tab)

The statistics table shows the results of the tests for each trait. The first line is for the model with no markers. Following tha is a single line for each marker tested. The columns labeled “df”, “F”, and “p” are the degrees of freedom, F, and p-value from the F distribution for the test of the marker. The column “errordf” is the degrees of freedom used for the denominator of the F-test. The column labeled “markerR2” is the R2 for the marker calculated based on a formula for R2 for a generalized least squares GLS) model as shown here.

**The columns “Genetic Var”, Residual Var”, and “-2LnLikelihood” list σ2a, σ2e, and minus two times the model likelihood, respectively. When the P3D option is used, all of the values are the same for a given trait because they are only calculated once. A second table lists the estimated effects of each allele for each marker similar to the output for GLM. The compression results table shown below shows the likelihood, genetic variance, and error variance for each compression level tested during the optimization process. The meaning of groups and compression is discussed above in the description of the compression method. The compression level with the lowest value of -2LnLk is used for testing markers.**


MLM_compression_for_Filtered_mdp_traits + Filtered_mdp_population_structure + mdp_...

| Trait | # groups | Compression | -2LnLk | Var_genetic | Var_error |
|---|---|---|---|---|---|
| dpoll | 259 | 1 | 1,480.402 | 7.362 | 8.146 |
| dpoll | 248 | 1.044 | 1,479.47 | 7.635 | 7.81 |
| dpoll | 243 | 1.066 | 1,479.505 | 7.656 | 7.855 |
| dpoll | 238 | 1.088 | 1,481.049 | 7.338 | 8.416 |
| dpoll | 234 | 1.107 | 1,482.935 | 6.957 | 9.069 |
| dpoll | 229 | 1.131 | 1,483.301 | 6.904 | 9.261 |
| dpoll | 224 | 1.156 | 1,482.597 | 6.866 | 9.394 |
| dpoll | 220 | 1.177 | 1,486.718 | 6.172 | 10.576 |
| dpoll | 215 | 1.205 | 1,485.526 | 6.407 | 10.342 |
| dpoll | 211 | 1.227 | 1,486.045 | 6.21 | 10.709 |
| dpoll | 207 | 1.251 | 1,488.214 | 5.897 | 11.345 |

Print Export (CSV) Export (Tab)

## 6.7 *Genomic Selection (using Ridge Regression)*

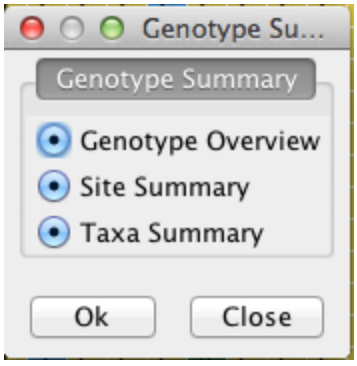
This function performs ridge regression to predict phenotypes from genotypes. It is one of the methods used for genomic selection (GS).

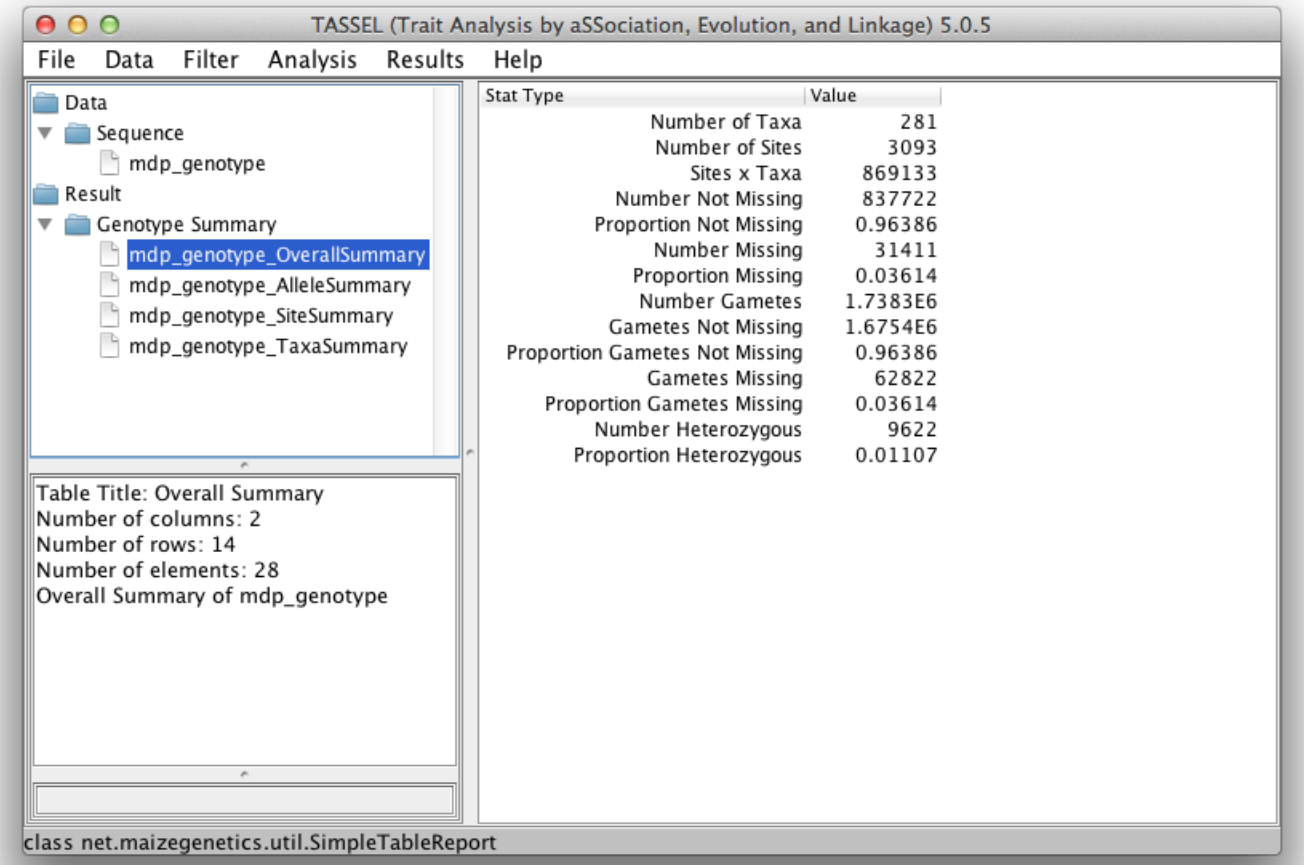
The input dataset must contain one or more phenotypes and numeric marker data. Optionally, it may also contain factors and covariates. The analysis is run by selecting the input dataset then clicking the "GS" button. Because no additional user input is needed, the analysis will run immediately after the button is clicked. All traits will be analyzed separately using all of the genotypes, factors, and covariates in the dataset. The output will consist of two new datasets for each trait. One of the datasets will contain genomic estimated breeding values (GEBVs) for each taxon and the other will contain BLUPs for each marker in the genotype file. The output datasets will appear in the "Numerical" folder, which holds the input data as well. The output datasets can in turn be used for subsequent analysis. For example, it could be joined with the input data so that the predicted values could be graphed against the original values.

Understanding the input data requirements is important to ensure that the results of the analysis will be correct and useful. Genotypes must be numeric with one column for each marker. It is expected that the markers are bi-allelic, with the homozygotes coded as 1 and -1 and the heterozygotes coded as 0. However, any reasonable coding scheme will work. For instance, missing data could be replaced by a probability resulting from imputation. If any genotype data is missing, it will be imputed as the average of the marker scores across all taxa for that marker. If a user prefers to use a different method of imputation, then the missing genotypes must be imputed before importing the data into TASSEL.

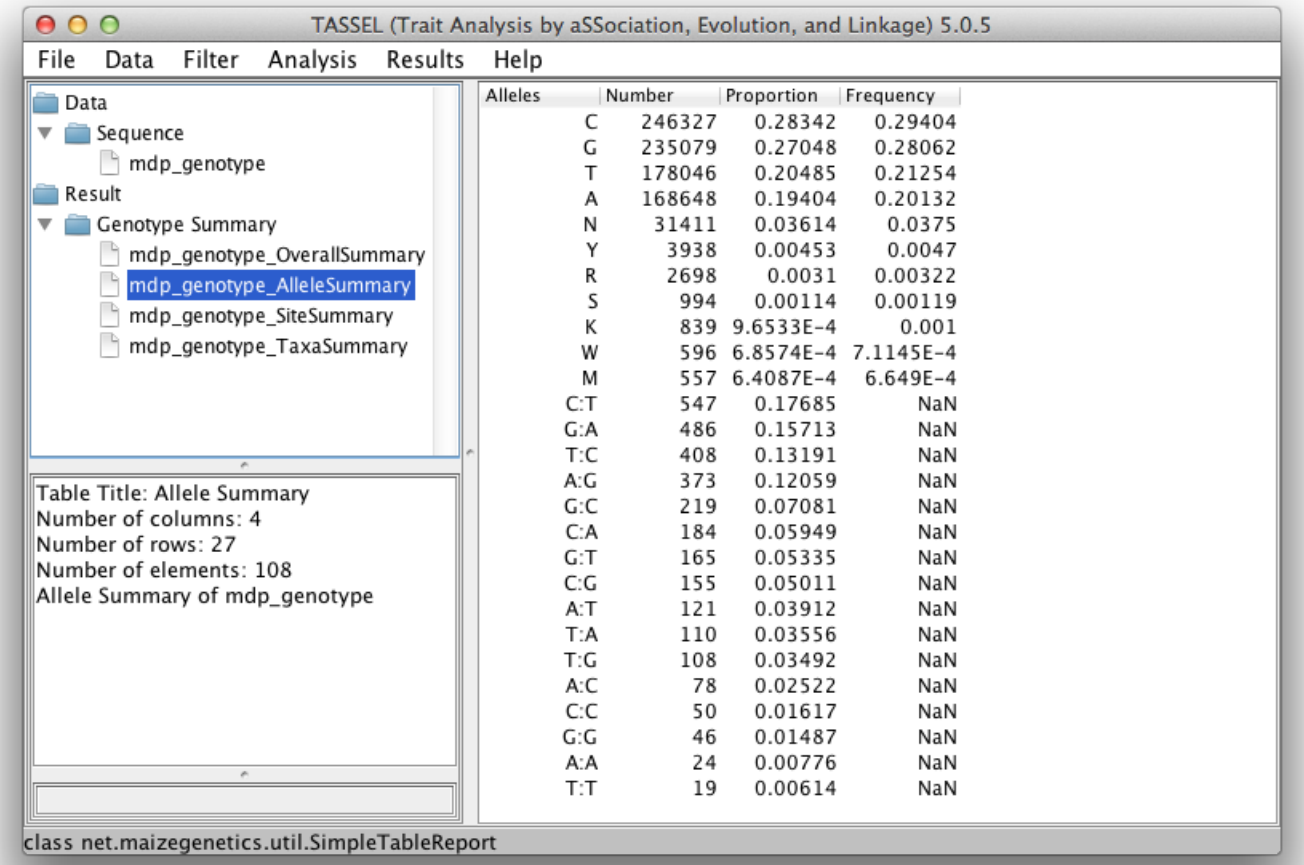
GEBVs will be calculated for all taxa in the dataset, including any lines that have missing phenotype data. A typical use of genomic selection is to predict GEBVs for a set of unphenotyped lines based on the performance of a training set. To do that a dataset containing both the genotypes to be predicted and the genotypes of the training set can be joined with a dataset containing the phenotypes of the training set using a union join. All taxa in the phenotype set should have genotypes. If an individual without genotype data is included, all the marker data for that individual will be imputed, which is not a generally useful thing to do.

## 6.8 *Geno Summary*

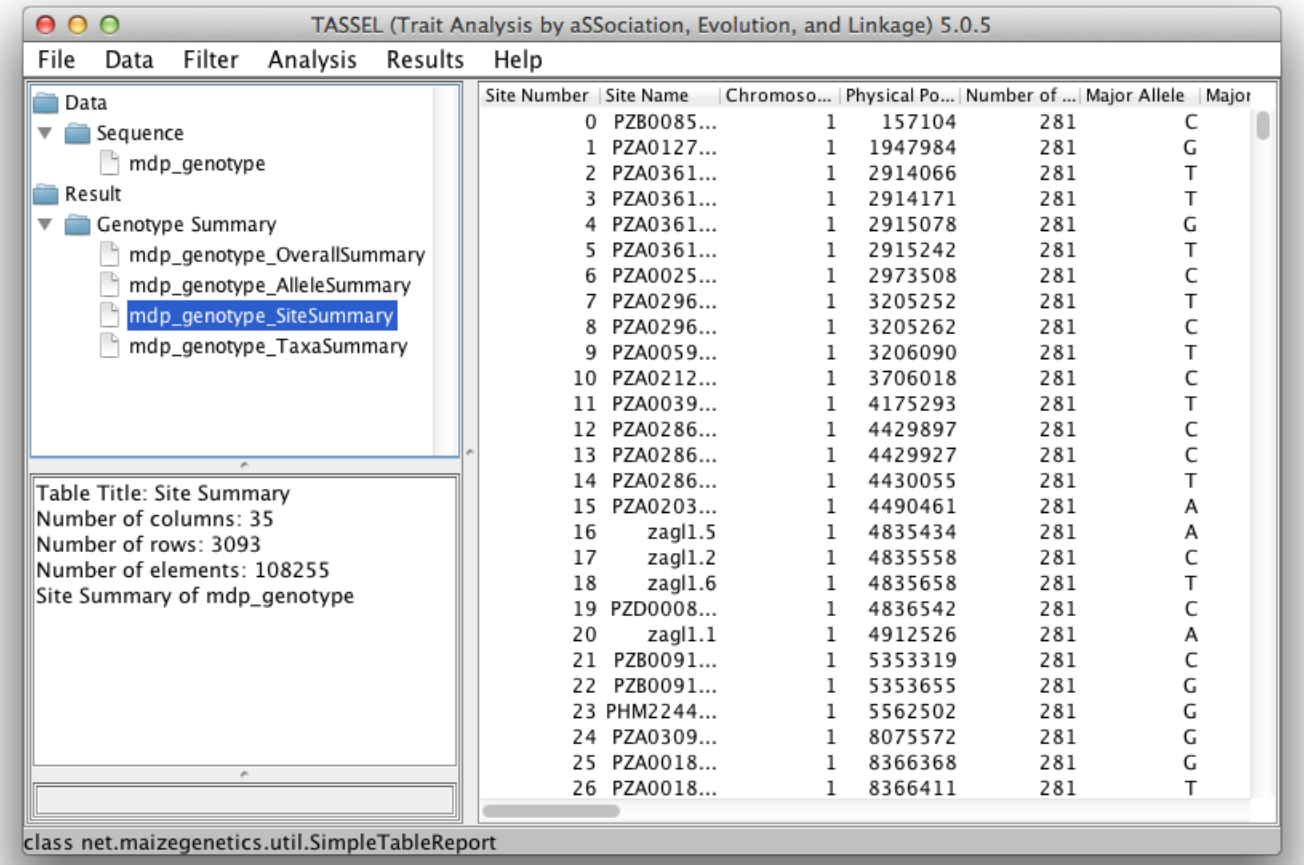





- Number of Taxa - Number of Taxa in data set.
- Number of Sites - Number of Sites in data set.
- Sites x Taxa - Number of sites multiplied by number of taxa.
- Number Not Missing - Number allele values not unknown (NN)
- Proportion Not Missing - Number Not Missing / Sites x Taxa
- Number Missing - Number unknown (NN) values
- Proportion Missing - Number Missing / Sites x Taxa
- Number Gametes - Number of Sites x Number Taxa x 2
- Gametes Not Missing - Number of gametes not unknown
- Proportion Gametes Not Missing - Gametes Not Missing / Number Gametes
- Gametes Missing - Number unknown (N) gametes
- Proportion Gametes Missing - Gametes Missing / Number Gametes
- Number Heterozygous - Number of heterozygous values
- Proportion Heterozygous - Number Heterozygous / Sites x Taxa

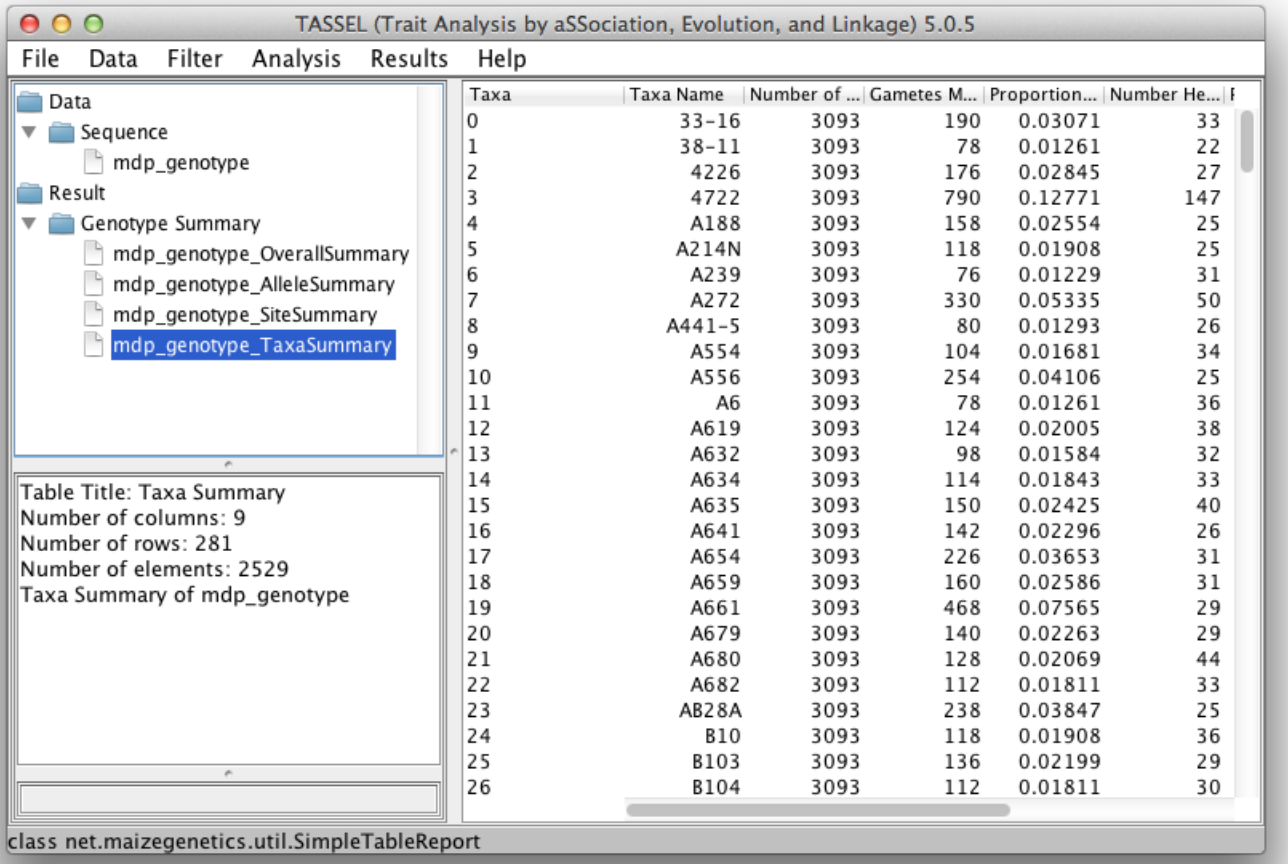



- Alleles - Allele values present in data set. Single letter values are diploid where some letter represent heterozygous. Two letter values are major / minor combinations with count of sites.
- Number - Number of occurrences
- Proportion - Percentage the value occurs in data set.
- Frequency - Percentage the value occurs in data set not counting unknown (N) values.


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.0.5
File Data Filter Analysis Results Help
Data
Sequence
mdp_genotype
Result
Genotype Summary
mdp_genotype_OverallSummary
mdp_genotype_AlleleSummary
mdp_genotype_SiteSummary
mdp_genotype_TaxaSummary
Table Title: Site Summary
Number of columns: 35
Number of rows: 3093
Number of elements: 108255
Site Summary of mdp_genotype
Site Number | Site Name | Chromoso... | Physical Po... | Number of ... | Major Allele | Major
0 PZB0085... 1 157104 281 C
1 PZA0127... 1 1947984 281 G
2 PZA0361... 1 2914066 281 T
3 PZA0361... 1 2914171 281 T
4 PZA0361... 1 2915078 281 G
5 PZA0361... 1 2915242 281 T
6 PZA0025... 1 2973508 281 C
7 PZA0296... 1 3205252 281 T
8 PZA0296... 1 3205262 281 C
9 PZA0059... 1 3206090 281 T
10 PZA0212... 1 3706018 281 C
11 PZA0039... 1 4175293 281 T
12 PZA0286... 1 4429897 281 C
13 PZA0286... 1 4429927 281 C
14 PZA0286... 1 4430055 281 T
15 PZA0203... 1 4490461 281 A
16 zagl1.5 1 4835434 281 A
17 zagl1.2 1 4835558 281 C
18 zagl1.6 1 4835658 281 T
19 PZD0008... 1 4836542 281 C
20 zagl1.1 1 4912526 281 A
21 PZB0091... 1 5353319 281 C
22 PZB0091... 1 5353655 281 G
23 PHM2244... 1 5562502 281 G
24 PZA0309... 1 8075572 281 G
25 PZA0018... 1 8366368 281 G
26 PZA0018... 1 8366411 281 T
class net.maizegenetics.util.SimpleTableReport

- Site Number - Index of site
- Site Name - Name of site
- Chromosome - Chromosome
- Physical Position - Physical Position on Chromosome
- Number of Taxa - Number of taxa for site (same of all)
- Major Allele - The major allele of site
- Major Allele Gametes - Number of times major allele occurs for site (up to twice number of taxa)
- Major Allele Proportion - Major Allele Gametes / (Number of Taxa * 2). Number of Taxa * 2 is the Number of Gametes for a Site.
- Major Allele Frequency - Major Allele Gametes / ((Number of Taxa * 2) - Gametes Missing)
- Minor Allele - The minor allele of site
- Minor Allele Gametes - Number of times minor allele occurs for site
- Minor Allele Proportion - Minor Allele Gametes / (Number of Taxa * 2). Number of Taxa * 2 is the Number of Gametes for a Site.
- Minor Allele Frequency - Minor Allele Gametes / ((Number of Taxa * 2) - Gametes Missing)
- Gametes Missing - Number of gametes with unknown (N) value
- Proportion Missing - Gametes Missing / (Number of Taxa * 2)
- Number Heterozygous - Number of taxa that are heterozygous for site.
- Proportion Heterozygous - Number Heterozygous / Number of Taxa (not counting taxa that are unknown (NN))
- Inbreeding Coefficient -

- Inbreeding Coefficient Scaled by Missing -




- Taxa - Index of taxa.
- Taxa Name - Name of taxa
- Number of Sites - Number of sites for taxon (same for all).
- Gametes Missing - Number of gametes with unknown (N) value. Every taxa / site combination has two gametes.
- Proportion Missing - Gametes Missing / (Number of Sites * 2)
- Number Heterozygous - Number of sites that are heterozygous for taxon
- Proportion Heterozygous - Number Heterozygous / Number of Sites (not counting sites that are unknown (NN))
- Inbreeding Coefficient -
- Inbreeding Coefficient Scaled by Missing -

## *6.9 Stepwise*

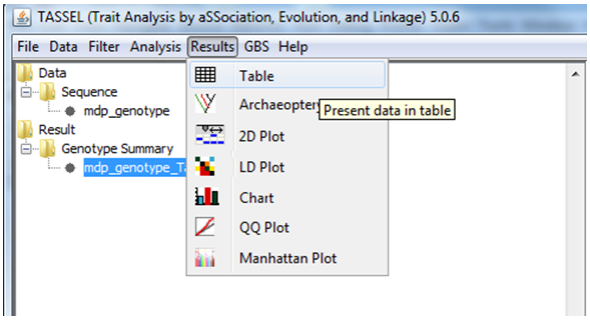
# 7 Results Menu

Results consists of the functions to present data as table or graphics.

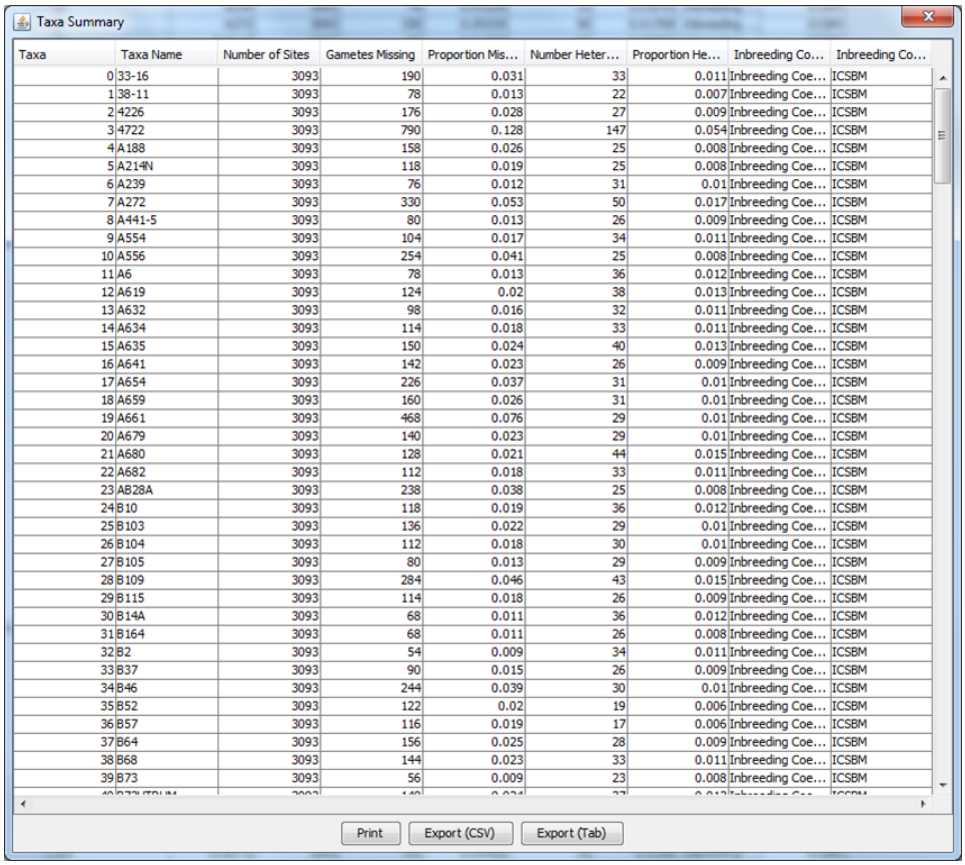
## *7.1 Table*

Allows data to be displayed in a spreadsheet view and exported into a flat file.

To create a table, select a data set from the Data Tree panel, then click on the menu “Results -> Table”.




Shown below is an example in which the Taxa Summary is displayed.



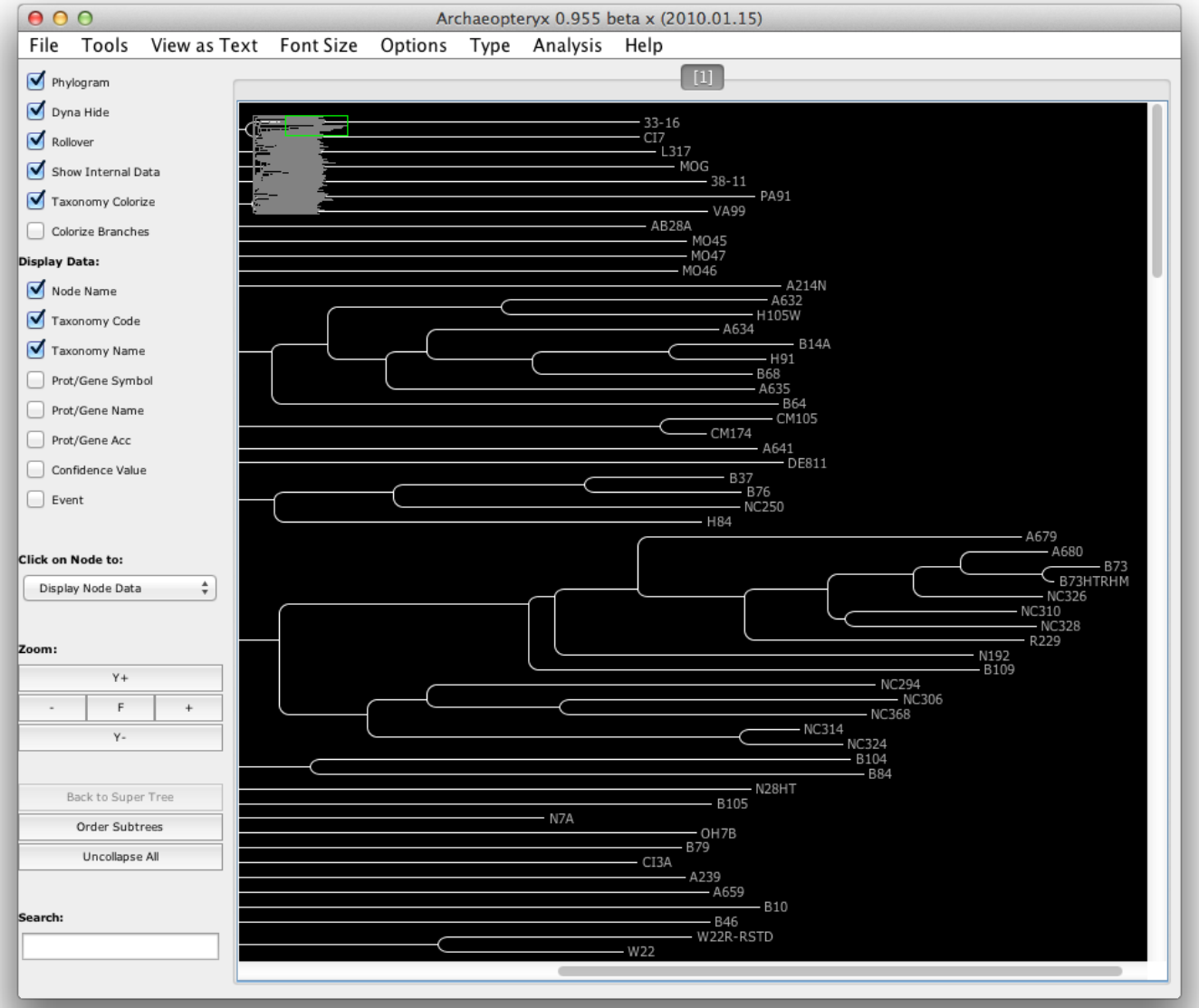

| Taxa | Taxa Name | Number of Sites | Gametes Missing | Proportion Mis... | Number Heter... | Proportion He... | Inbreeding Co... | Inbreeding Co... |
|---|---|---|---|---|---|---|---|---|
| 0 | 33-16 | 3093 | 190 | 0.031 | 33 | 0.011 | Inbreeding Coe... | ICSBM |
| 1 | 38-11 | 3093 | 78 | 0.013 | 22 | 0.007 | Inbreeding Coe... | ICSBM |
| 2 | 4226 | 3093 | 176 | 0.028 | 27 | 0.009 | Inbreeding Coe... | ICSBM |
| 3 | 4722 | 3093 | 790 | 0.128 | 147 | 0.054 | Inbreeding Coe... | ICSBM |
| 4 | A188 | 3093 | 158 | 0.026 | 25 | 0.008 | Inbreeding Coe... | ICSBM |
| 5 | A214N | 3093 | 118 | 0.019 | 25 | 0.008 | Inbreeding Coe... | ICSBM |
| 6 | A239 | 3093 | 76 | 0.012 | 31 | 0.01 | Inbreeding Coe... | ICSBM |
| 7 | A272 | 3093 | 330 | 0.053 | 50 | 0.017 | Inbreeding Coe... | ICSBM |
| 8 | A441-5 | 3093 | 80 | 0.013 | 26 | 0.009 | Inbreeding Coe... | ICSBM |
| 9 | A554 | 3093 | 104 | 0.017 | 34 | 0.011 | Inbreeding Coe... | ICSBM |
| 10 | A556 | 3093 | 254 | 0.041 | 25 | 0.008 | Inbreeding Coe... | ICSBM |
| 11 | A6 | 3093 | 78 | 0.013 | 36 | 0.012 | Inbreeding Coe... | ICSBM |
| 12 | A619 | 3093 | 124 | 0.02 | 38 | 0.013 | Inbreeding Coe... | ICSBM |
| 13 | A632 | 3093 | 98 | 0.016 | 32 | 0.011 | Inbreeding Coe... | ICSBM |
| 14 | A634 | 3093 | 114 | 0.018 | 33 | 0.011 | Inbreeding Coe... | ICSBM |
| 15 | A635 | 3093 | 150 | 0.024 | 40 | 0.013 | Inbreeding Coe... | ICSBM |
| 16 | A641 | 3093 | 142 | 0.023 | 26 | 0.009 | Inbreeding Coe... | ICSBM |
| 17 | A654 | 3093 | 226 | 0.037 | 31 | 0.01 | Inbreeding Coe... | ICSBM |
| 18 | A659 | 3093 | 160 | 0.026 | 31 | 0.01 | Inbreeding Coe... | ICSBM |
| 19 | A661 | 3093 | 468 | 0.076 | 29 | 0.01 | Inbreeding Coe... | ICSBM |
| 20 | A679 | 3093 | 140 | 0.023 | 29 | 0.01 | Inbreeding Coe... | ICSBM |
| 21 | A680 | 3093 | 128 | 0.021 | 44 | 0.015 | Inbreeding Coe... | ICSBM |
| 22 | A682 | 3093 | 112 | 0.018 | 33 | 0.011 | Inbreeding Coe... | ICSBM |
| 23 | AB28A | 3093 | 238 | 0.038 | 25 | 0.008 | Inbreeding Coe... | ICSBM |
| 24 | B10 | 3093 | 118 | 0.019 | 36 | 0.012 | Inbreeding Coe... | ICSBM |
| 25 | B103 | 3093 | 136 | 0.022 | 29 | 0.01 | Inbreeding Coe... | ICSBM |
| 26 | B104 | 3093 | 112 | 0.018 | 30 | 0.01 | Inbreeding Coe... | ICSBM |
| 27 | B105 | 3093 | 80 | 0.013 | 29 | 0.009 | Inbreeding Coe... | ICSBM |
| 28 | B109 | 3093 | 284 | 0.046 | 43 | 0.015 | Inbreeding Coe... | ICSBM |
| 29 | B115 | 3093 | 114 | 0.018 | 26 | 0.009 | Inbreeding Coe... | ICSBM |
| 30 | B14A | 3093 | 68 | 0.011 | 36 | 0.012 | Inbreeding Coe... | ICSBM |
| 31 | B164 | 3093 | 68 | 0.011 | 26 | 0.008 | Inbreeding Coe... | ICSBM |
| 32 | B2 | 3093 | 54 | 0.009 | 34 | 0.011 | Inbreeding Coe... | ICSBM |
| 33 | B37 | 3093 | 90 | 0.015 | 26 | 0.009 | Inbreeding Coe... | ICSBM |
| 34 | B46 | 3093 | 244 | 0.039 | 30 | 0.01 | Inbreeding Coe... | ICSBM |
| 35 | B52 | 3093 | 122 | 0.02 | 19 | 0.006 | Inbreeding Coe... | ICSBM |
| 36 | B57 | 3093 | 116 | 0.019 | 17 | 0.006 | Inbreeding Coe... | ICSBM |
| 37 | B64 | 3093 | 156 | 0.025 | 28 | 0.009 | Inbreeding Coe... | ICSBM |
| 38 | B68 | 3093 | 144 | 0.023 | 33 | 0.011 | Inbreeding Coe... | ICSBM |
| 39 | B73 | 3093 | 56 | 0.009 | 23 | 0.008 | Inbreeding Coe... | ICSBM |

Data can be sorted by clicking on the column header of interest. A secondary sort can be done by holding down the CTRL key and clicking on a second column.

Data can be exported to flat files that are either comma-separated (Comma Separated Values = CSV) or tab-delimited. Both these formats can then be imported into a spreadsheet program such as Excel. Tables can also be printed.

## *7.2 Archaeopteryx Tree*

Select a “Tree:...” data set to use. https://sites.google.com/site/cmzmasek/home/software/archaeopteryx

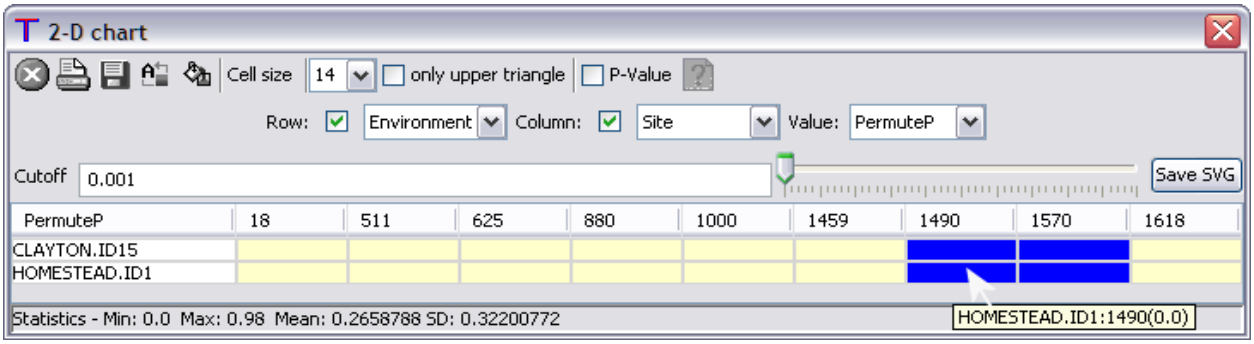



## 7.3 2D Plot

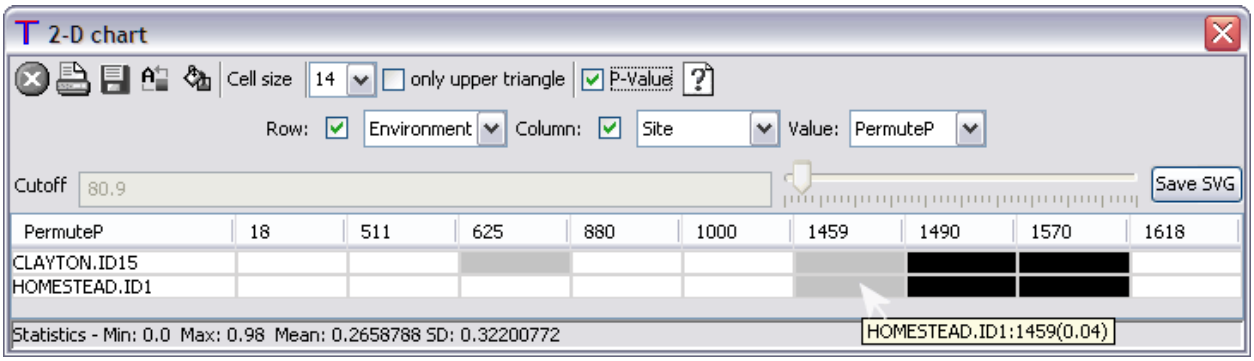
Displays 2D plots and determines color thresholds.

This function is useful for plotting associations in multiple environments.

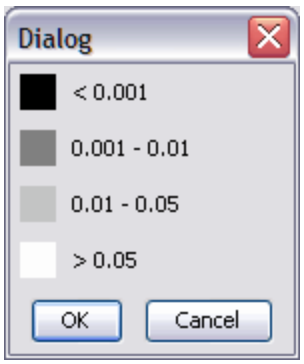
First, select the desired result set. Using the drop down boxes provided, populate rows with “Environment,” columns with “Site,” and value with “PermuteP.” The cutoff value for coloring can be chosen either by inputting a value in the te box or by using the slider tool to the right of the text box. Users can “mouse over” any box to view the value associated with that box, as shown here:




If P-value coloring is desired, simply check the P-value box as shown below:




By checking the P-value box, Cutoff selection tools will be disabled and fields will instead be colored according to the following grayscale:




This key can be shown by clicking on the "?" icon next to the P-value check box.

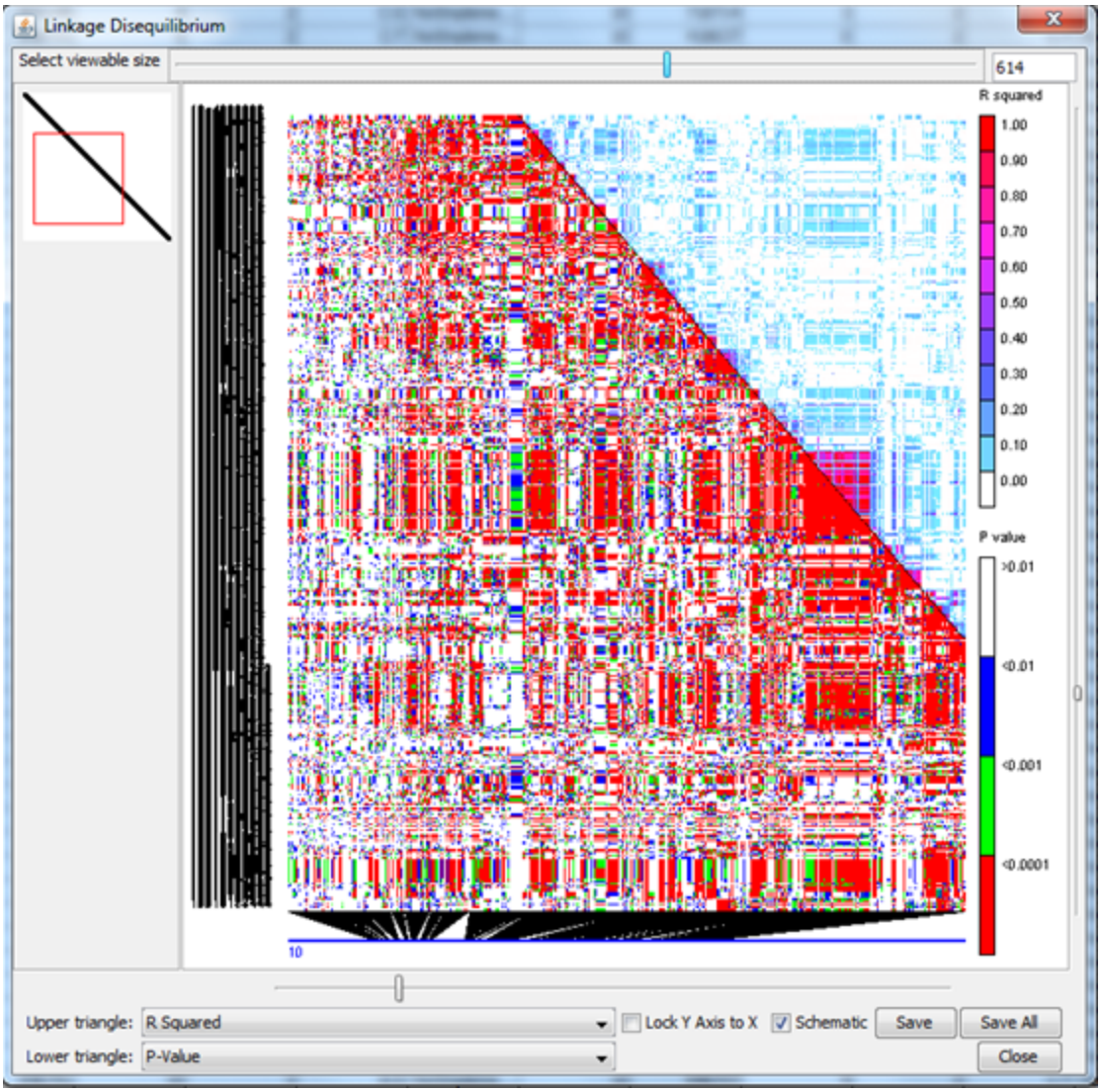
## *7.4 LD Plot*

Displays the results from a linkage disequilibrium analysis.

After selecting the desired result from the Data Tree, choose Results -> LD Plot

The graph that is generated displays LD between pairs of sites calculated with the analysis step. The black diagonal represents LD between each site and itself. The default setting graphs $r^2$ in the upper right and *p*-values in the lower left

This default can be modified by clicking on the buttons in the lower left. The left side of the graph contains a tex description with the Chromosome and the Site name. At the bottom of the graph is a display of the position of each site along the chromosome. This display can be hidden by deselecting the “Schematic” check box. Legends that describe the color scheme appear on the right hand side of the graph.
The number of sites displayed can be selected by entering a number in the white box in the upper right corner or by moving the sliding bar next to it. To move through the graph use the sliding bars on the right and bottom. The red box i the small white window in the upper left corner will show what portion of the graph is displayed. To move only aroun the diagonal select the “Lock Y Axis to X” check box (recommended when visualizing a LD by sliding window analysis).



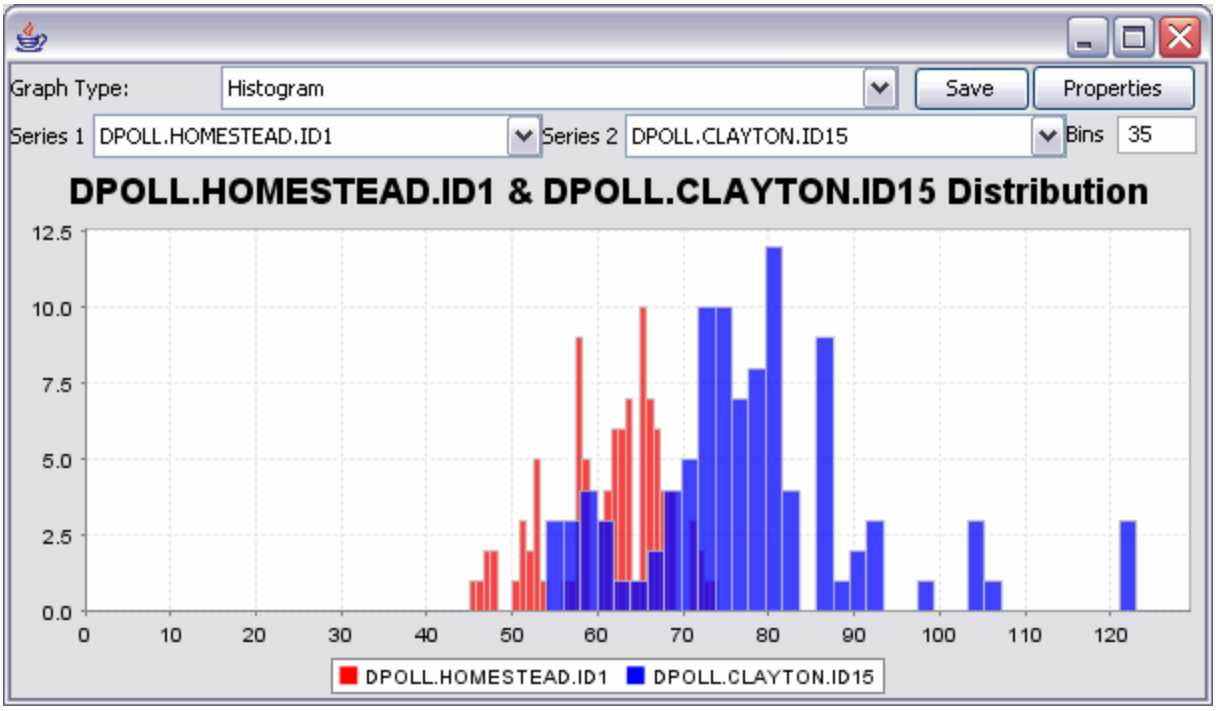

LD plots can be saved in several formats. The Save button will save the area of the graph shown in the screen, while the Save All button will save the entire graph.

## 7.5 Chart

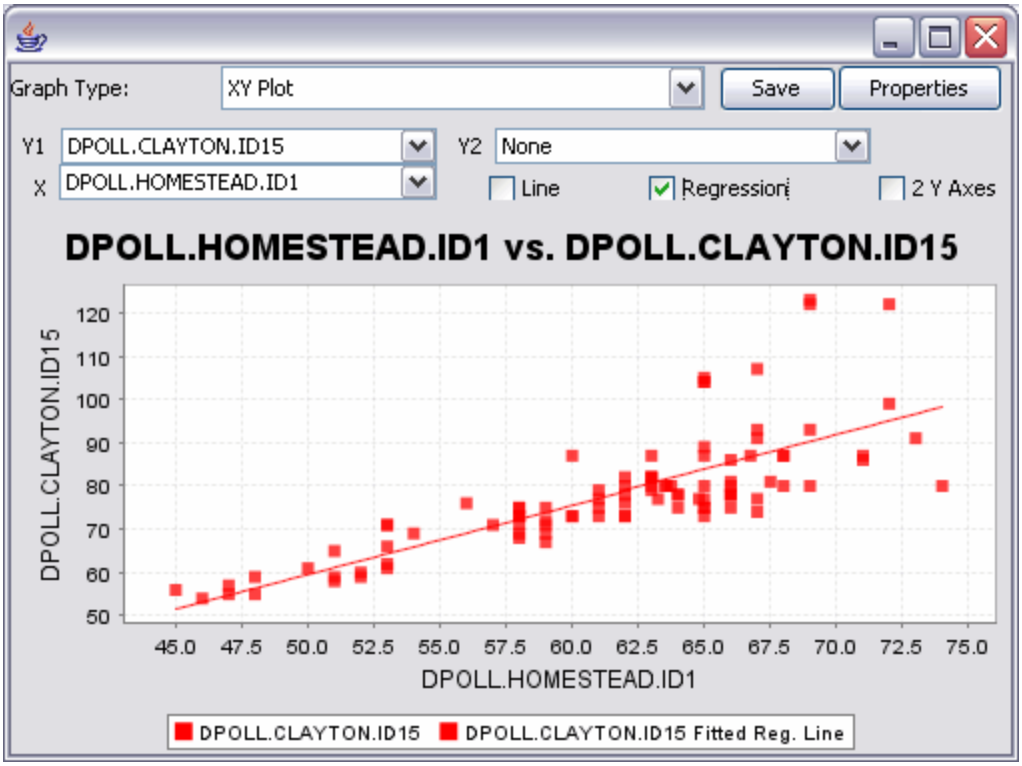
Chart provides a variety of graphs for visualizing numeric data.

This feature can be used to display histograms, XY plots, bar charts and/or pie charts. Any numeric table data can be charted, including LD results, phenotypic data, diversity results, and association results.

Histograms: Use the graph type combo box to select the desired graph type (Histogram) from the list of options. Up to two different series of data can be plotted together. Users may specify the number of bins to be used in the histogram.




Scatter plots: Use the graph type combo box to select the desired graph type (XY Plot) from the list of options. Select data to be plotted in X and Y axes using the appropriate drop down boxes. If two data series are plotted simultaneously on the Y axis, the “2 Y Axes” checkbox will provide an axis for each.




### *7.6 QQ Plot*

### *7.7 Manhattan Plot*

## 8 GBS Menu

http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf

## 9 Help Menu

Help provides information Tassel and diagnostics.

### *9.1 Help Manual*

### *9.2 About*

### *9.3 Show Memory*

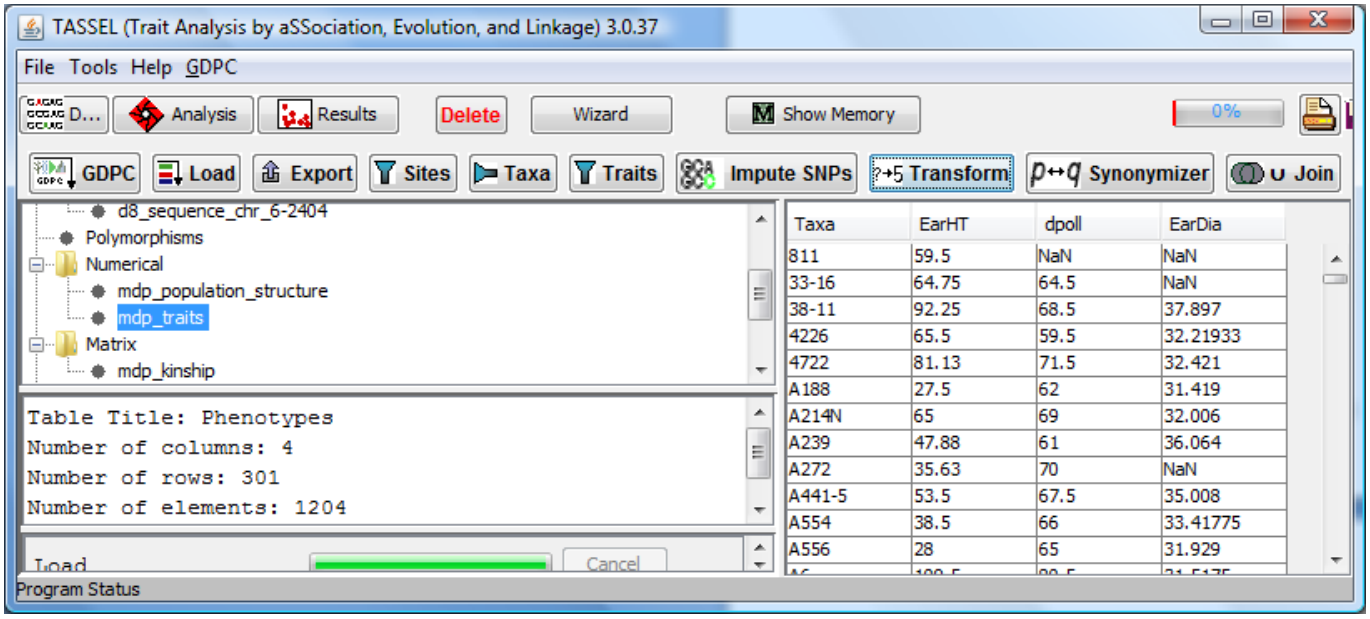
### *9.4 Logging*

# 10 Tutorial

This tutorial reviews several common scenarios for using TASSEL in order to help the user better understand its capabilities for data manipulation and association analyses. The TASSEL software package includes a tutorial data set that can be downloaded from the TASSEL website (please unzip all files to a directory of your choice). This tutorial data set contains data for phenotype, genotype, population structure, and kinship.

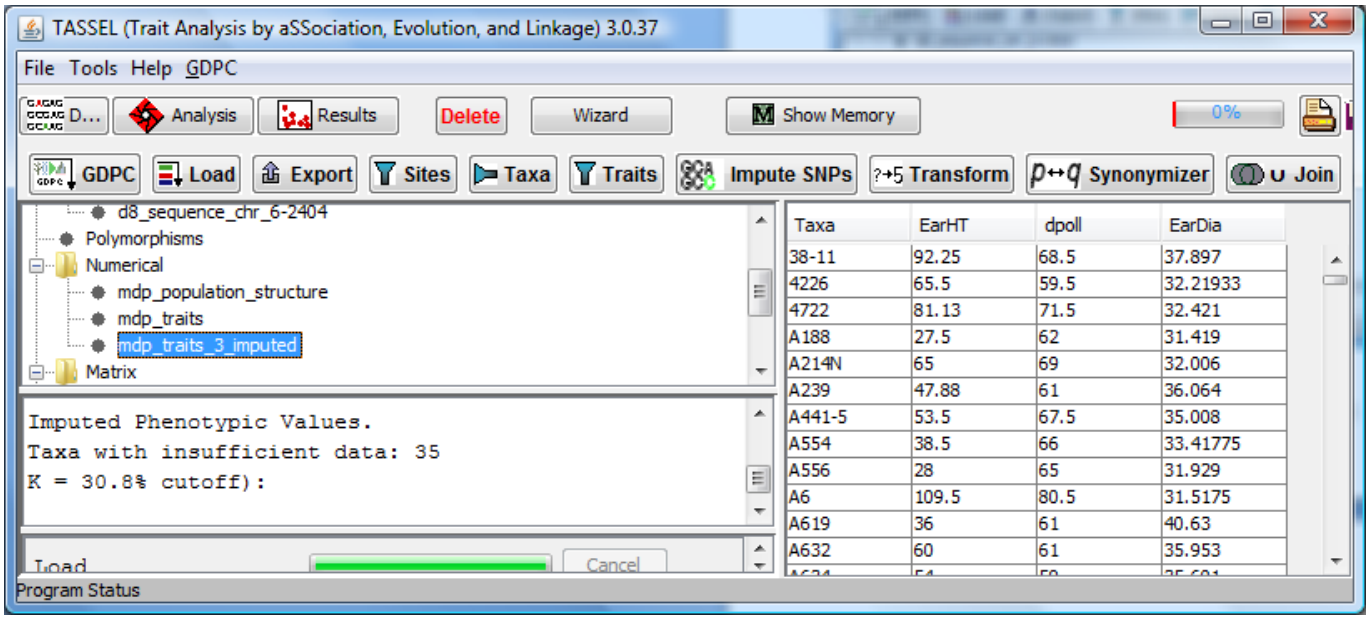
## *10.1 Missing Phenotype Imputation*

The phenotype file **mdp_traits** will be used to demonstrate the process of imputing missing data. Note that the data set below contains missing values (NaN).




To impute missing data, first select the **mdp_traits** data set in the Data Tree Panel and then click the **Transform** button (**Data □ Transform**). The "Transform Column Data" window will open. Click on the **Impute** tab in this window. Finally, click on the **Create Data set** button to create the new data set with missing values imputed.

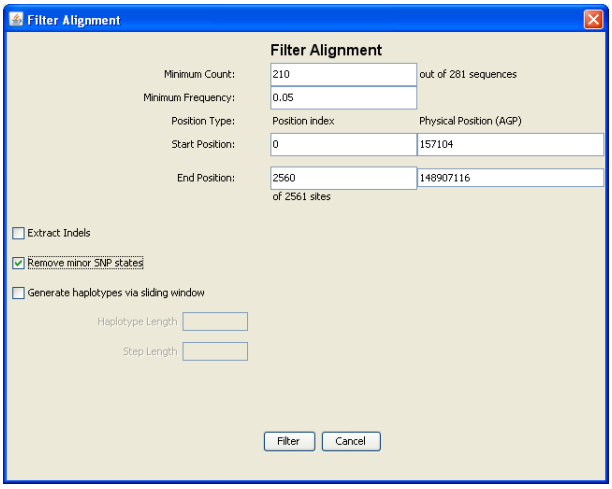
Note that missing values are now filled.

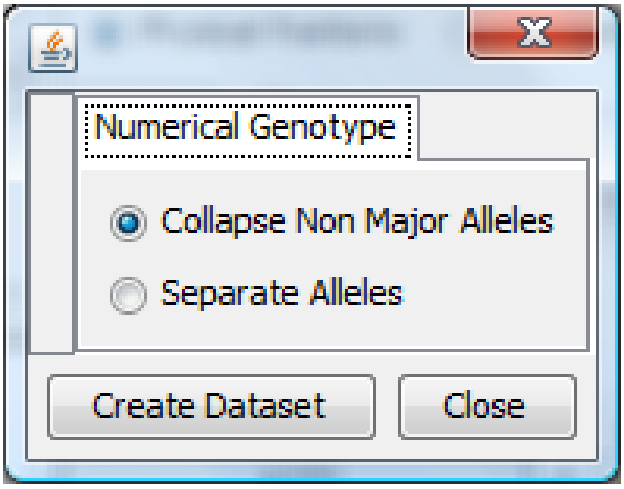


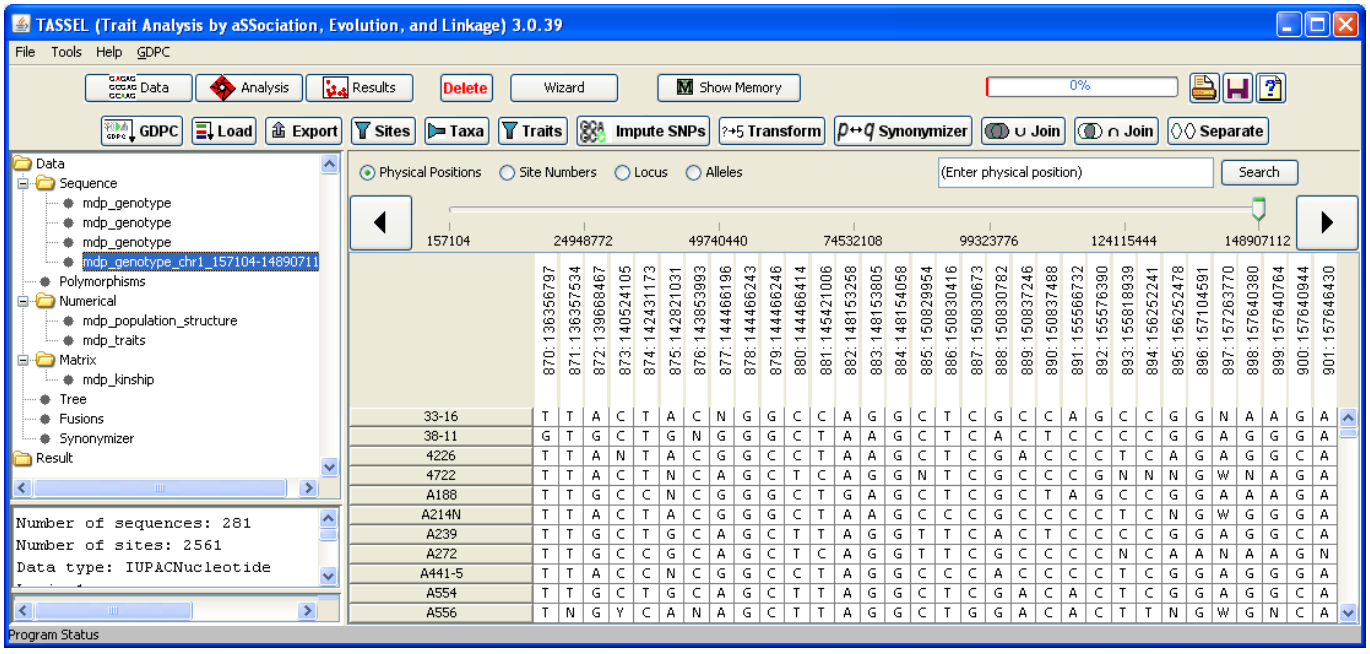

## *10.2 Principal Component Analysis*

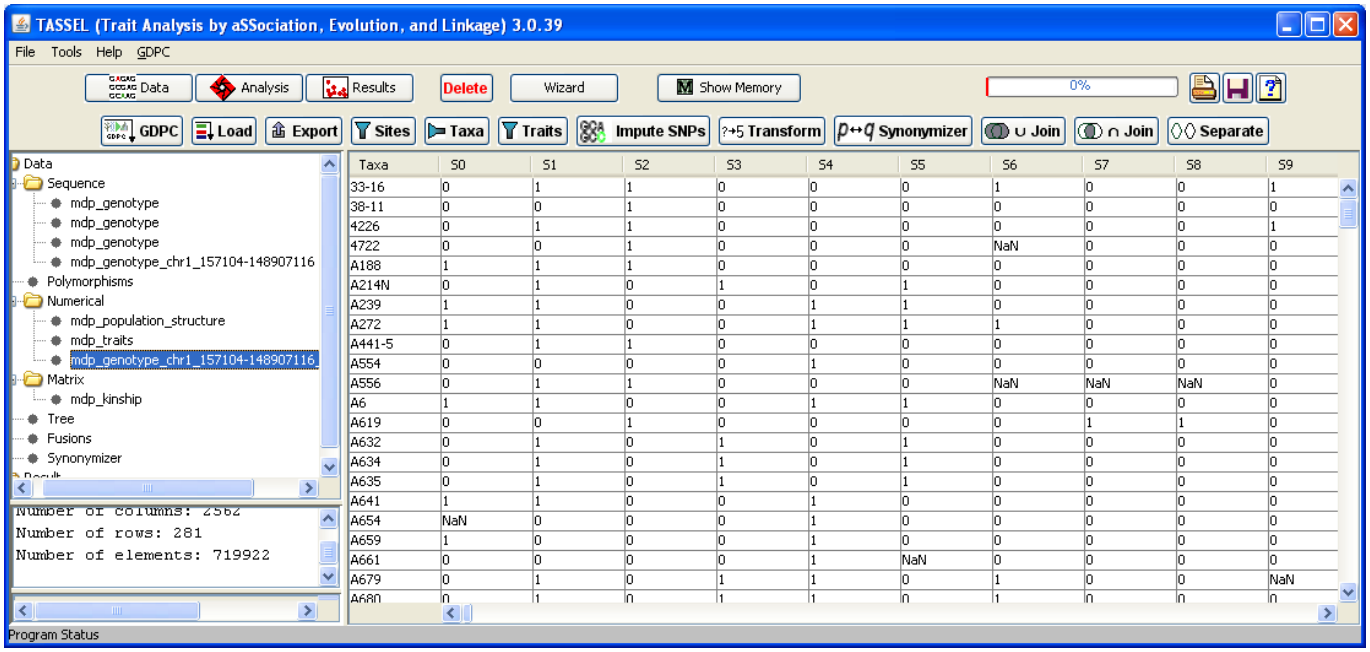
Principal component analysis (PCA) is a statistical tool that transforms a set of correlated variables into a smaller number of uncorrelated variables called principal components (PCs). The first PC captures as much of the variation as possible, and the succeeding PCs account for a decreasing fraction of the remaining variance. Another application of PCA is to use PCs derived from genetic markers to represent population structure[8]. This method requires much less computing time than maximum likelihood estimation. As most marker data are characters, numericalization must be performed first. A common approach for converting character marker scores is to set one of the homozygotes to 0, the other homozygote to 2, and the heterozygote to 1. For haploids, the conversion can be simply performed by coding one allele as 0 and the other as 1. The TRANSFORM function in TASSEL converts the major allele to 0. All the other alleles are collapsed to a single class and coded as 1. PCA requires that all variables should have variation and should not have missing values. As a result, filtering genotype to eliminate monomorphic markers and imputing missing values may be necessary. Imputing missing values can be done before or after numericalization. Here we demonstrate how to generate PCs from the genotype file in the tutorial data.

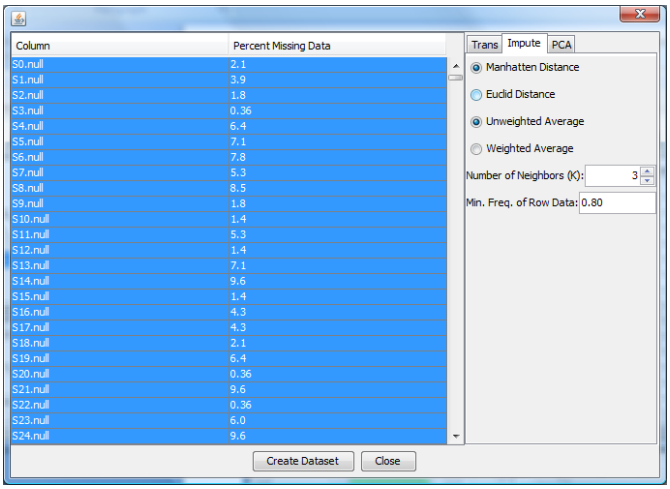
1. Remove monomorphic sites: Make sure TASSEL is in **Data** mode. Highlight the genotype and click **Site**. Set the minimum frequency to 0.05 and have "Remove minor SNP status" checked. Click **Filter**.
2. Numericalization: Highlight the filtered genotype and click **Transform**. Use the default option of "Collapse non maj alleles." Click **Create data set**.
3. Imputation of missing values: Highlight the numerical genotype and click **Transform** and then click **Impute** Tab. U the default options. Click **Create data set**.
4. PCA: Highlight the imputed numerical genotype, click **Transform,** and then click **PCA** Tab. Change the default optio to "Components=3" by choosing **Components** and type 3 in the text box. Click **Create data set**.

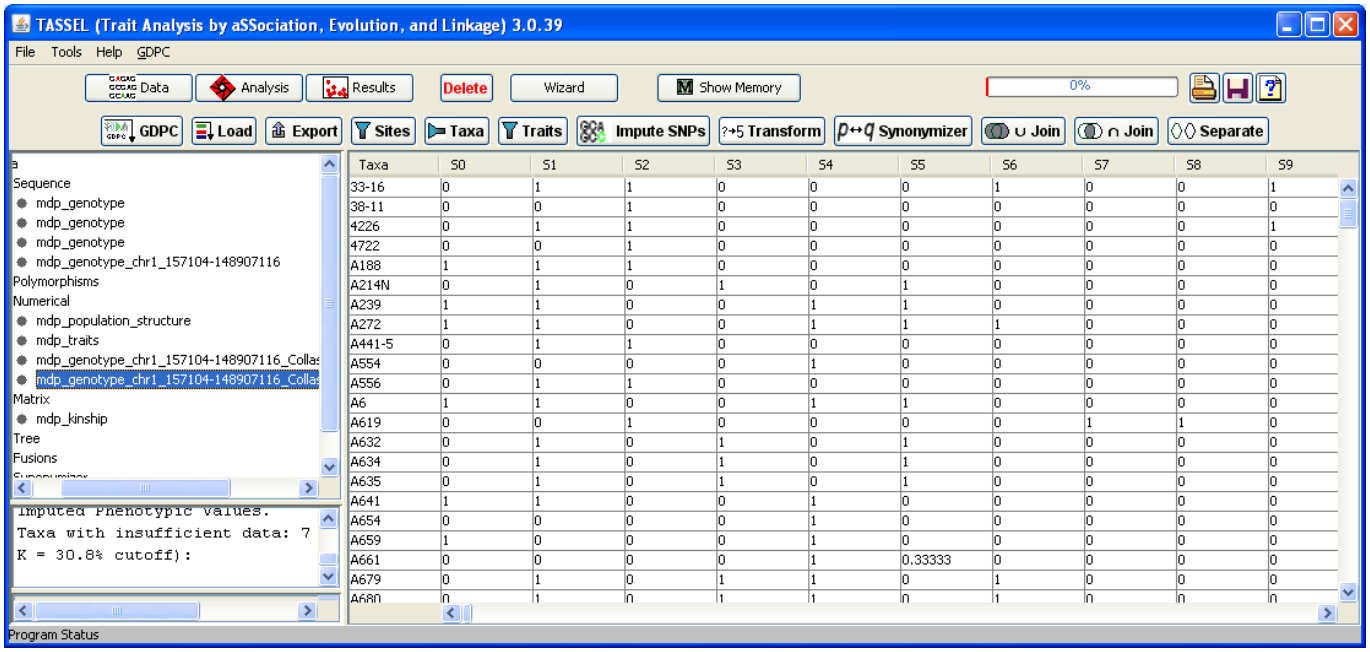

Filter Alignment
Filter Alignment
Minimum Count:
210
out of 281 sequences
Minimum Frequency:
0.05
Position Type:
Position index
Physical Position (AGP)
Start Position:
0
157104
End Position:
2560
148907116
of 2561 sites
Extract Indels
Remove minor SNP states
Generate haplotypes via sliding window
Haplotype Length
Step Length
Filter
Cancel

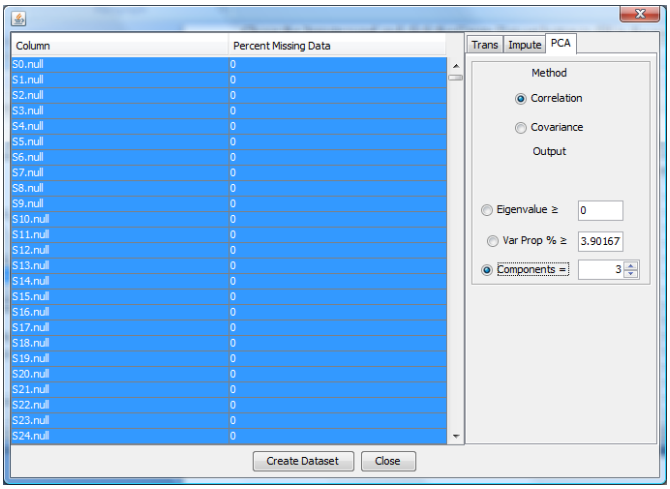

Numerical Genotype
Collapse Non Major Alleles
Separate Alleles
Create Dataset
Close


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File
Tools
Help
GDPC
Data
Analysis
Results
Delete
Wizard
Show Memory
0%
GDPC
Load
Export
Sites
Taxa
Traits
Impute SNPs
Transform
Synonymizer
Join
Join
Separate
Data
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-14890711
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
Matrix
mdp_kinship
Tree
Fusions
Synonymizer
Result
Physical Positions
Site Numbers
Locus
Alleles
(Enter physical position)
Search
157104
24948772
49740440
74532108
99323776
124115444
148907112
870:136356797
871:136357534
872:139668467
873:140524105
874:142431173
875:142821031
876:143853993
877:144466196
878:144466243
879:144466246
880:144466414
881:145421006
882:148153258
883:148153805
884:148154058
885:150829954
886:150830416
887:150830673
888:150830782
889:150837246
890:150837488
891:155566732
892:155576390
893:155818939
894:156252241
895:156252478
896:157104591
897:157263770
898:157640380
899:157640764
900:157640944
901:157646430
33-16 T T A C T A C N G G C C A G G C T C G C C A G C C G G N A A G A
38-11 G T G C T G N G G G C T A A G C T C A C T C C C C G G A G G G A
4226 T T A N T A C G G C C T A A G C T C G A C C C T C A G A G G C A
4722 T T A C T N C A G C C T C A G G N T C G C C C C G N N N G W N A G A
A188 T T G C C N C G G G C T G A G C T C G C T A G C C G G A A A G A
A214N T T A C T A C G G G C T A A G C C C G C C C T C N G W G G G A
A239 T T G C T G C A G C T T A G G T T C A C T C C C C G G A G G C A
A272 T T G C C G C A G C T C A G G T T C G C C C C N C A A N A A G N
A441-5 T T A C C N C G G C C T A G G C C C C A C C C C T C G G A G G G A
A554 T T G C T G C A G C T T A G G C T C G A C A C T C G G A G G C A
A556 T N G Y C A N A G C T T A G G C T C G A C A C T T N G W G N C A
Number of sequences: 281
Number of sites: 2561
Data type: IUPACNucleotide
Program Status

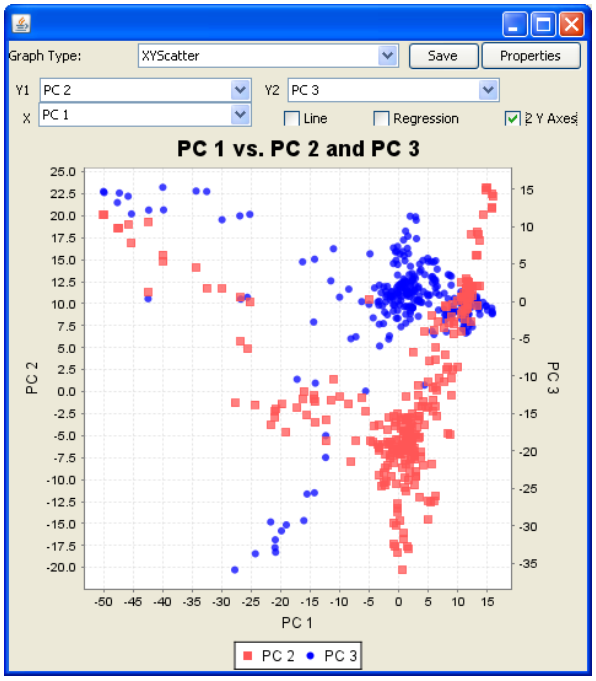

TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File Tools Help GDPC
Data Analysis Results Delete Wizard Show Memory 0%
GDPC Load Export Sites Taxa Traits Impute SNPs ?→5 Transform p↔q Synonymizer ∪ Join ∩ Join Separate
Data
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-148907116
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
mdp_genotype_chr1_157104-148907116
Matrix
mdp_kinship
Tree
Fusions
Synonymizer
Number of columns: 2562
Number of rows: 281
Number of elements: 719922
Taxa S0 S1 S2 S3 S4 S5 S6 S7 S8 S9
33-16 0 1 1 0 0 0 1 0 0 1
38-11 0 0 1 0 0 0 0 0 0 0
4226 0 1 1 0 0 0 0 0 0 1
4722 0 0 1 0 0 0 NaN 0 0 0
A188 1 1 1 0 0 0 0 0 0 0
A214N 0 1 0 1 0 1 0 0 0 0
A239 1 1 0 0 1 1 0 0 0 0
A272 1 1 0 0 1 1 1 0 0 0
A441-5 0 1 1 0 0 0 0 0 0 0
A554 0 0 0 0 1 0 0 0 0 0
A556 0 1 1 0 0 0 NaN NaN NaN 0
A6 1 1 0 0 1 1 0 0 0 0
A619 0 0 1 0 0 0 0 1 1 0
A632 0 1 0 1 0 1 0 0 0 0
A634 0 1 0 1 0 1 0 0 0 0
A635 0 1 0 1 0 1 0 0 0 0
A641 1 1 0 0 1 0 0 0 0 0
A654 NaN 0 0 0 1 0 0 0 0 0
A659 1 0 0 0 1 0 0 0 0 0
A661 0 0 0 0 1 NaN 0 0 0 0
A679 0 1 0 1 1 0 1 0 0 NaN
Program Status

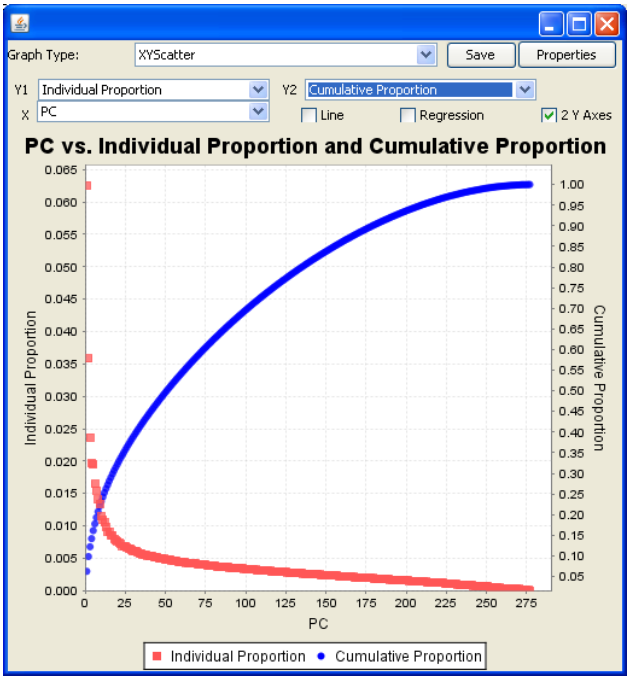

Column Percent Missing Data
S0.null 2.1
S1.null 3.9
S2.null 1.8
S3.null 0.36
S4.null 6.4
S5.null 7.1
S6.null 7.8
S7.null 5.3
S8.null 8.5
S9.null 1.8
S10.null 1.4
S11.null 5.3
S12.null 1.4
S13.null 7.1
S14.null 9.6
S15.null 1.4
S16.null 4.3
S17.null 4.3
S18.null 2.1
S19.null 6.4
S20.null 0.36
S21.null 9.6
S22.null 0.36
S23.null 6.0
S24.null 9.6
Trans Impute PCA
Manhatten Distance
Euclid Distance
Unweighted Average
Weighted Average
Number of Neighbors (K): 3
Min. Freq. of Row Data: 0.80
Create Dataset
Close


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 3.0.39
File Tools Help GDPC
Data Analysis Results Delete Wizard Show Memory 0%
GDPC Load Export Sites Taxa Traits Impute SNPs ?→5 Transform p↔q Synonymizer ∪ Join ∩ Join Separate
Sequence
mdp_genotype
mdp_genotype
mdp_genotype
mdp_genotype_chr1_157104-148907116
Polymorphisms
Numerical
mdp_population_structure
mdp_traits
mdp_genotype_chr1_157104-148907116_Colla
mdp_genotype_chr1_157104-148907116_Colla
Matrix
mdp_kinship
Tree
Fusions
Imputed Phenotypic Values.
Taxa with insufficient data: 7
K = 30.8% cutoff):
Program Status
Taxa S0 S1 S2 S3 S4 S5 S6 S7 S8 S9
33-16 0 1 1 0 0 0 1 0 0 1
38-11 0 0 1 0 0 0 0 0 0 0
4226 0 1 1 0 0 0 0 0 0 1
4722 0 0 1 0 0 0 0 0 0 0
A188 1 1 1 0 0 0 0 0 0 0
A214N 0 1 0 1 0 1 0 0 0 0
A239 1 1 0 0 1 1 0 0 0 0
A272 1 1 0 0 1 1 1 0 0 0
A441-5 0 1 1 0 0 0 0 0 0 0
A554 0 0 0 0 1 0 0 0 0 0
A556 0 1 1 0 0 0 0 0 0 0
A6 1 1 0 0 1 1 0 0 0 0
A619 0 0 1 0 0 0 0 1 1 0
A632 0 1 0 1 0 1 0 0 0 0
A634 0 1 0 1 0 1 0 0 0 0
A635 0 1 0 1 0 1 0 0 0 0
A641 1 1 0 0 1 0 0 0 0 0
A654 0 0 0 0 1 0 0 0 0 0
A659 1 0 0 0 1 0 0 0 0 0
A661 0 0 0 0 1 0.33333 0 0 0 0
A679 0 1 0 1 1 0 1 0 0 0




Three items will be added to the data tree after running PCA. The first are the PCs. The second are the eigenvalues. And, the last are the eigenvectors. Here we use the Chart Function in the Result mode to graph the first three PCs, the individual eigenvalue contributions (sometimes called a skree plot) and the cumulative eigenvalue contributions. T eigenvalues are of interest because they equal the variance explained by each of the PCs.

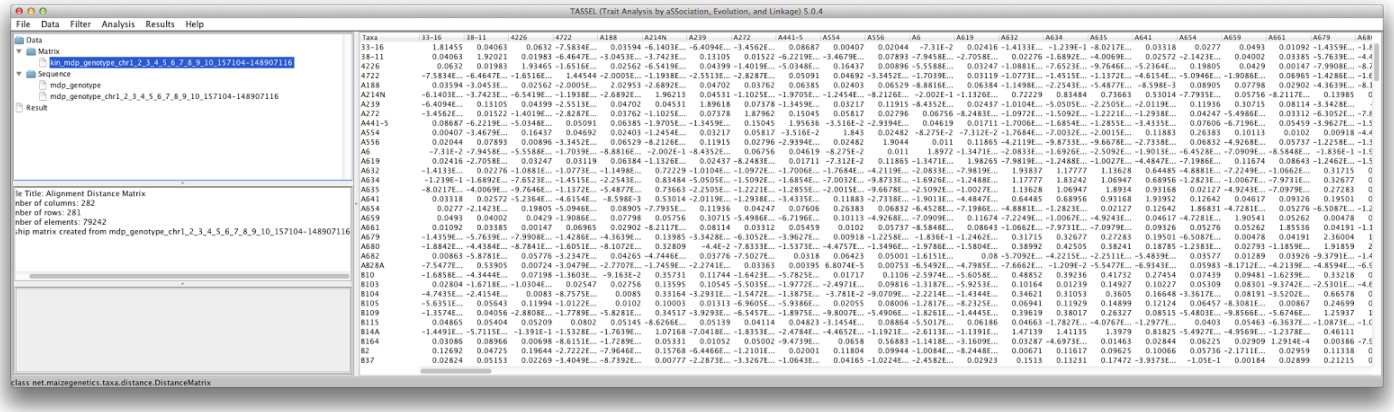



## *10.3 Estimation of Kinship using genetic markers*

While PCs can be used to capture major population subdivisions, kinship can be used to capture more subtle relationships. This section shows how to create a kinship matrix based on the same SNP data used to calculate PC's.

1. Remove monomorphic sites: Highlight the genotype and choose **Filter/Sites** on the menu bar. Set the threshold on MAF to 0.05, check "Remove minor SNP status," then click **Filter**.
2. Estimate kinship: Highlight the filtered genotype and click **Analysis/Kinship.** Leave "Scaled IBS" selected in th "Choose Kinship Method" dialog and click **OK**. A kinship matrix will be added to the data tree under Matrix category.
3. Alternatively, impute missing genotype data first then create the kinship matrix using the imputed data. To impute miss data, highlight the filtered genotype, choose Data/Transform, leave "Collapse Non-Major Alleles" selected, and click "Create Dataset". A new data set with "_Collapse" appended will appear in the "Numerical" folder. Highlight the collapsed data set, choose Data/Transform, select the Impute tab, then click "Create Dataset". Highlight the resulting imputed data then choose Analysis/Kinship.

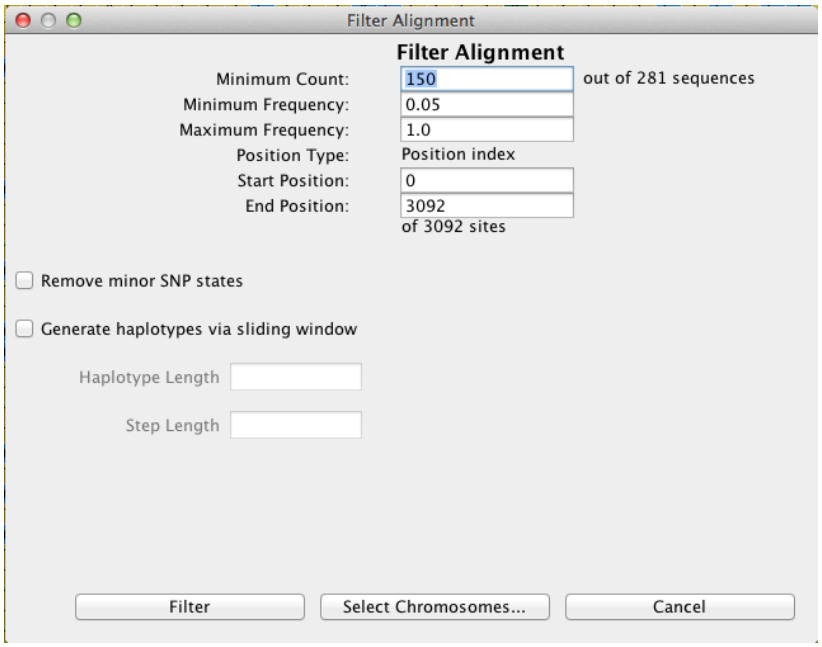


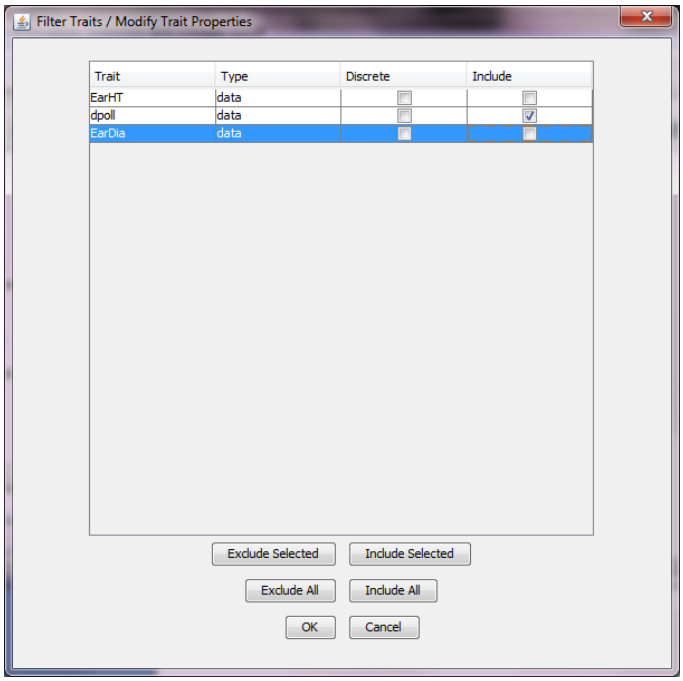

## *10.4 Association analysis using GLM*

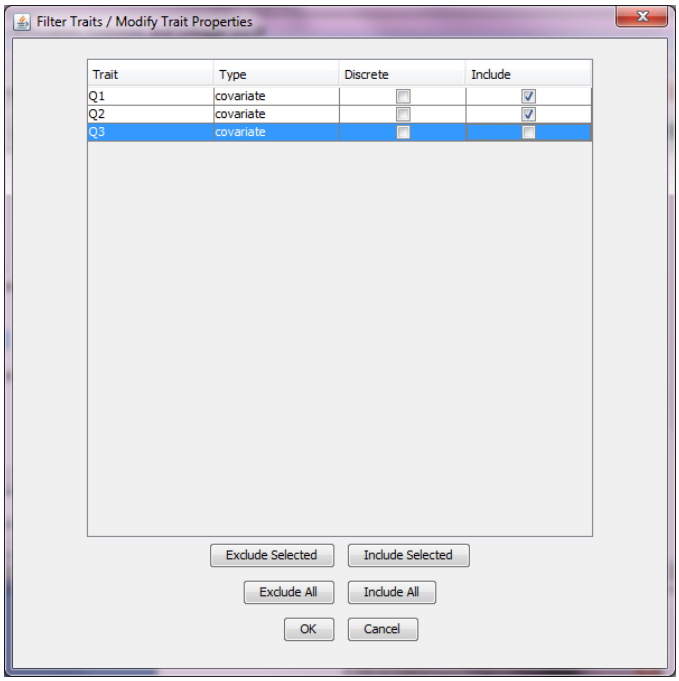
We use three files from the tutorial data set to perform association analysis using the **GLM**. The first file mdp_genotype.hmp.txt, a set of SNPs scored at 3093 sites on 281 maize inbred lines. The second one is the population structure of 282 maize inbred lines (mdp_population_structure.txt). The last one is phenotypes for three traits, for 282 maize inbred lines (mdp_traits.txt). The statistical model is:

Flowering time = Population structure + Marker effect + residual

1. Remove monomorphic and low coverage sites: Highlight the mdp_genotype and click **Filter/Sites** on the menu bar. Set "Minimum Frequency" to 0.05, "Maximum Frequency"to 1.0, and "Minimum Count" to 150. Click **Filter** to create a filtered genotype data set.
2. Trait selection: Highlight the phenotype and click the menu item **Filter/Traits**. Uncheck all the traits except floweri time (DPOLL). Make sure that the Type is set to Data. Click **OK** to create a filtered phenotype.
3. Covariate selection: The population structure is presented as the proportion of each population. There are three populations represented as Q1, Q2, and Q3. They sum to 100%. This creates linear dependency if we use all of them as covariates. While GLM can handle that properly, it will cause MLM to complain and refuse to complete your analysis. We can eliminate the dependency by removing one of the Q variables. In this demonstration, we exclude the last one. Highlight mdp_population_structure and click **Filter/Traits**. Uncheck the last population (Q3). Make sure that the Type is set to Covariate. Then click **OK** to create a filtered population structure data.
4. Joining data: Highlight the three filtered data sets by holding the Control key while selecting the individual data sets. Th click the menu item **Data/Intersect Join** to create a combined data set.
5. Association analysis: Highlight the joint data set then click the menu item **Analysis/GLM** to perform associatio analysis. Two reports will be added to the data tree.

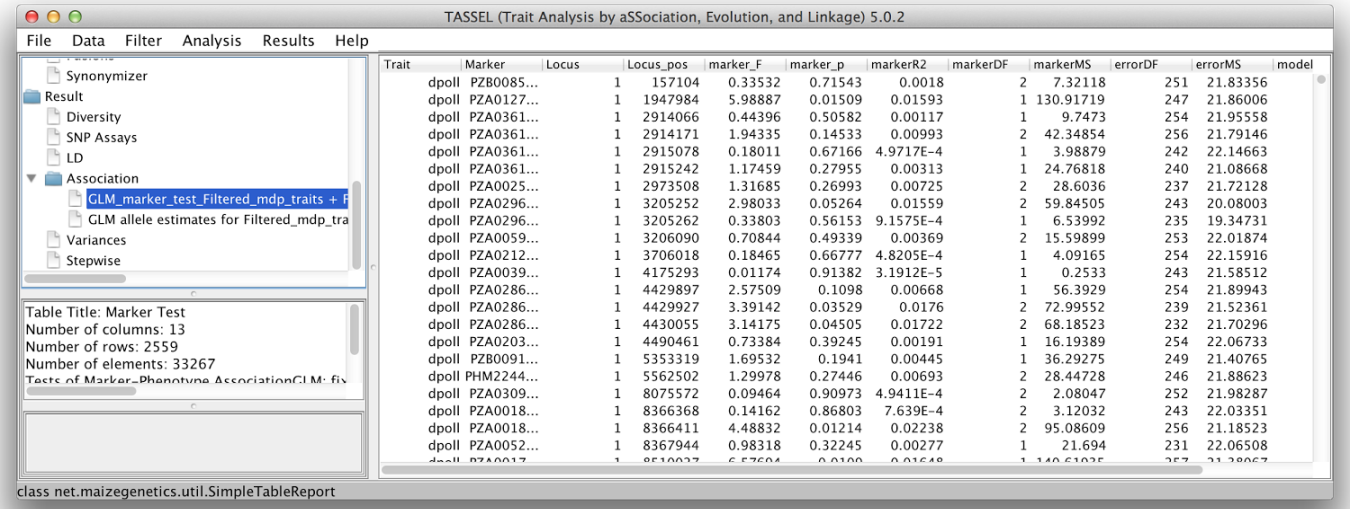


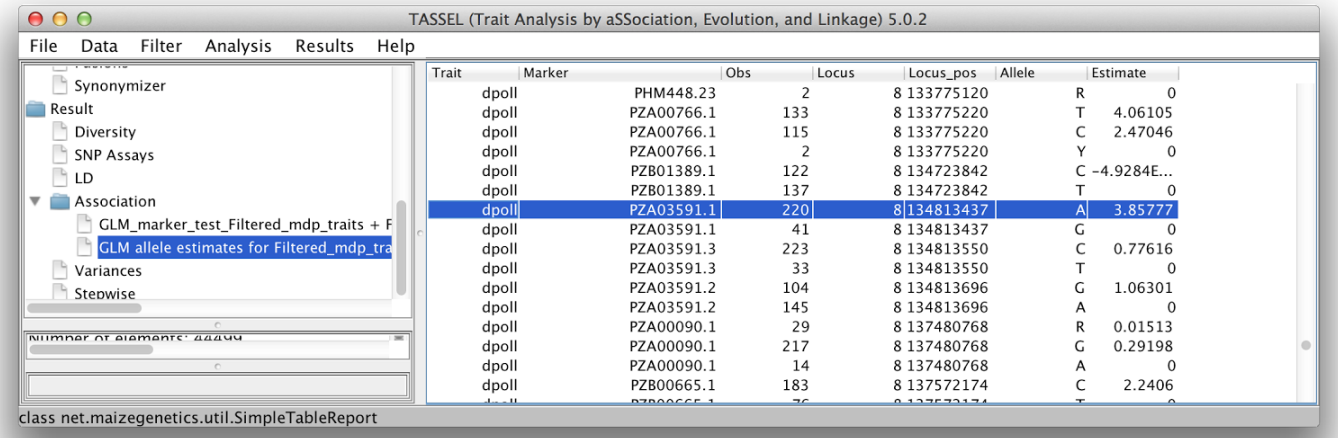

One of the reports added to data tree is labeled "GLM_Marker_Test_" followed by the name of the joint data. In addition to the information for traits and markers, the data set contains the following statistics:

marker_F: F value from the F test on marker;
marker_p: P value from the F test on marker;
markerR2: $R^2$ for the marker after fitting other model terms (population structure);
markerDF: Degree freedom of marker;
markerMS: Mean square of marker;
errorDF: Degree freedom of residual error;
errorMS: Mean square of residual error;
modelDF: Degree freedom of model;
modelMS: Mean square of model.




Clicking "marker_p" will sort the table by P value. The smallest P value is $3.5963 \times 10^{-6}$. A reasonable significance threshold is $1.9 \times 10^{-5}$, which is 5% after Bonferroni multiple test correction (0.05/2559). The denominator in the Bonferroni correction is the total number of SNPs tested. The association was significant.

The other data added to the data tree is labeled "GLM_Allele_Estimates_" followed by the name of the joint data. For the most significant SNP (highlighted in the figure below), there were two genotypes (AA and GG). There are 220 lines with genotype AA and 41 lines with allele GG. For the trait dpoll (days to pollination), the difference between the two homozygotes was 3.86 days.


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.0.2
File Data Filter Analysis Results Help
Synonymizer
Result
Diversity
SNP Assays
LD
Association
GLM_marker_test_Filtered_mdp_traits + F
GLM allele estimates for Filtered_mdp_tra
Variances
Stepwise
Trait Marker Obs Locus Locus_pos Allele Estimate
dpoll PHM448.23 2 8 133775120 R 0
dpoll PZA00766.1 133 8 133775220 T 4.06105
dpoll PZA00766.1 115 8 133775220 C 2.47046
dpoll PZA00766.1 2 8 133775220 Y 0
dpoll PZB01389.1 122 8 134723842 C -4.9284E...
dpoll PZB01389.1 137 8 134723842 T 0
dpoll PZA03591.1 220 8 134813437 A 3.85777
dpoll PZA03591.1 41 8 134813437 G 0
dpoll PZA03591.3 223 8 134813550 C 0.77616
dpoll PZA03591.3 33 8 134813550 T 0
dpoll PZA03591.2 104 8 134813696 G 1.06301
dpoll PZA03591.2 145 8 134813696 A 0
dpoll PZA00090.1 29 8 137480768 R 0.01513
dpoll PZA00090.1 217 8 137480768 G 0.29198
dpoll PZA00090.1 14 8 137480768 A 0
dpoll PZB00665.1 183 8 137572174 C 2.2406
class net.maizegenetics.util.SimpleTableReport

## *10.5 Association analysis using MLM*

Running MLM in tassel is similar to running GLM. The difference is that in addition to the joint data (or numerical data MLM requires kinship data to define the relationship between individuals. The kinship matrix times a parameter equal

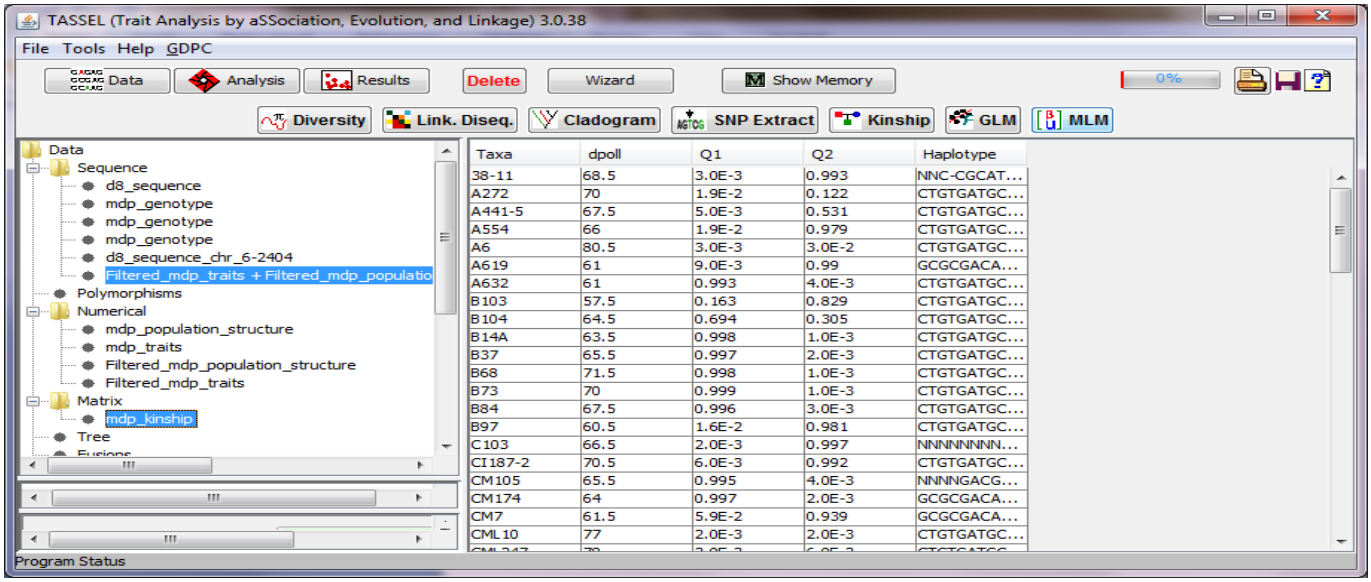
the covariance matrix between individuals. Here we use kinship file from the tutorial data set to fit the following statisti model.

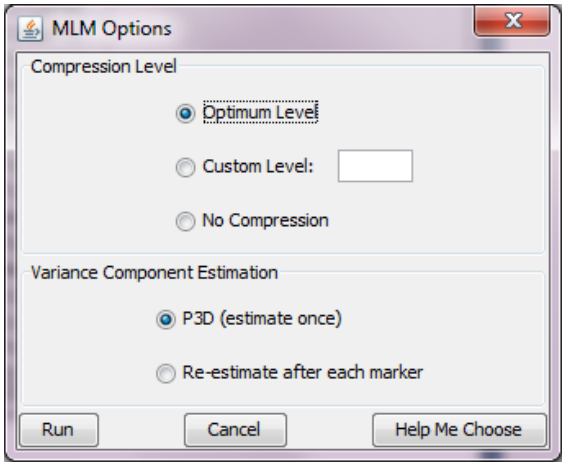
Flowering time = Population structure + Marker effect + Individuals + residual

Individuals and the residual are fit as random effects. The other terms are treated as fixed effects.

With respect to the marker effect, we will demonstrate the analysis using two sets of markers. One is the dwarf8 gene sequence used in the GLM tutorial. The other is a set of 3093 SNPs spread across the maize genome.

For the dwarf8 gene sequence, use the joint data set created by following the tutorial for GLM. Solve the mixed linear model by highlighting the joint data set and the kinship data then clicking the **MLM** button in **Analysis** mode.

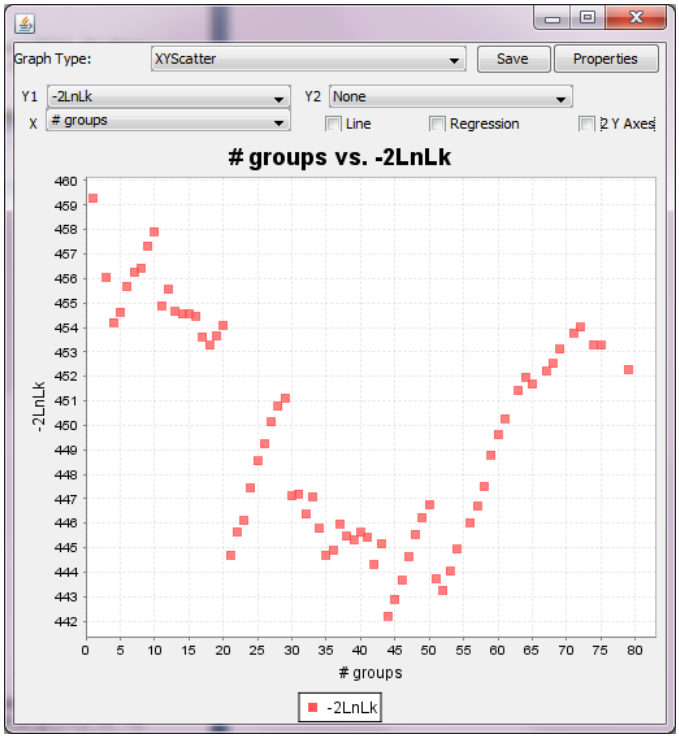


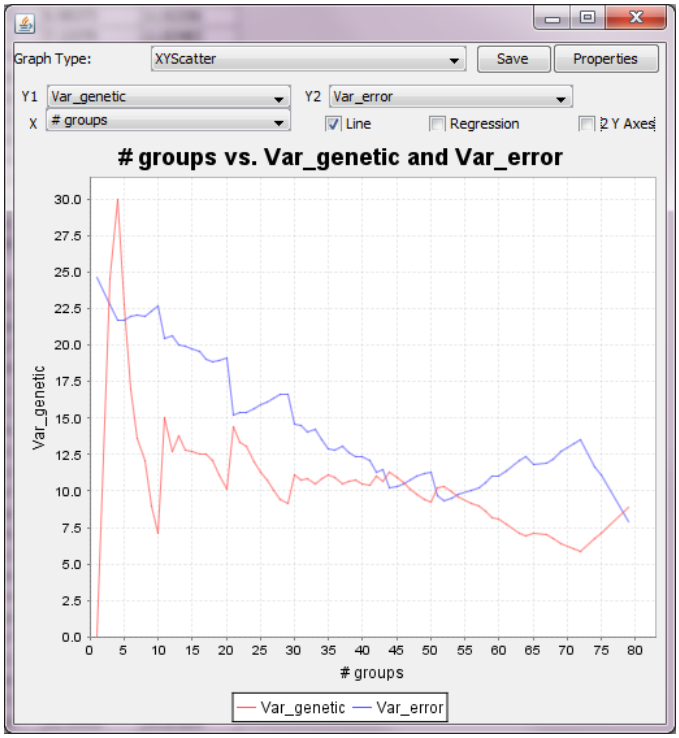




An MLM option dialog will pop up as shown above. Choose the default options, which use P3D and compression at the optimum compression level. After the Run button is clicked, the progress bar will start moving. The time require will depend on sample size, number of traits, number of markers, and the options chosen in the MLM option dialog. After the progress bar is reset to zero, indicating completion of MLM, three reports will be added to the data tree. The first two are similar to the reports created by GLM. The most significant SNP is still the same, however the strength o association is weaker, with a P value of $7.199 \times 10^{-4}$ (vs. $1.1021 \times 10^{-4}$ from GLM) which does not pass the Bonferroni

multiple test threshold ($5x10^{-4}$).

The third report contains the MLM specific statistics, including -2 Log Likelihood, genetic variance and residua variance components under different level of compression. These statistics are illustrated by the Chart function on th Result mode as follows.

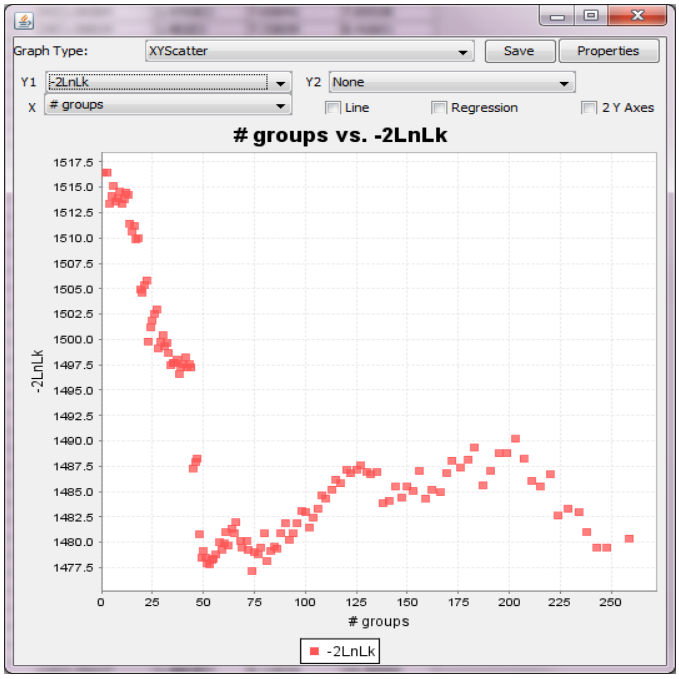


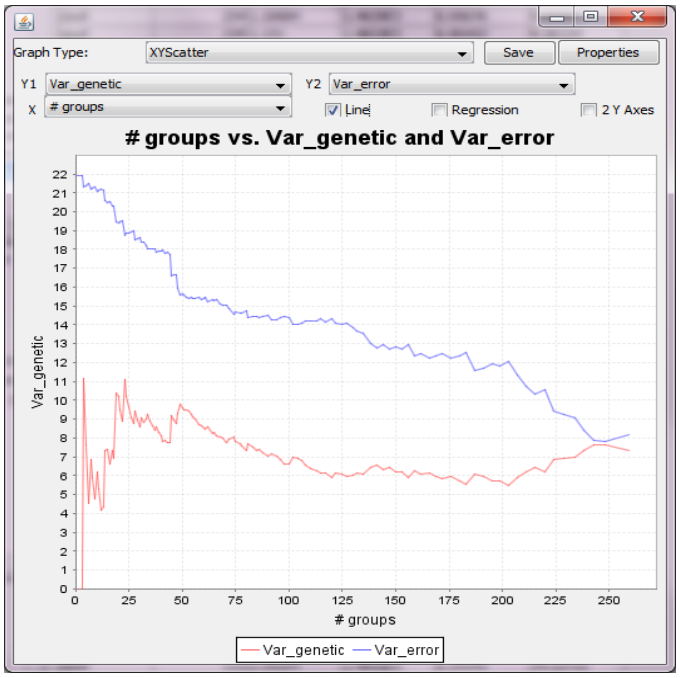




In the example, 79 are included in the final analysis. When they are clustered into 44 groups, the -2 Log Likelihood reaches a minimum, which indicates the best model fit. The screening of SNPs was performed at this optimum compression level.

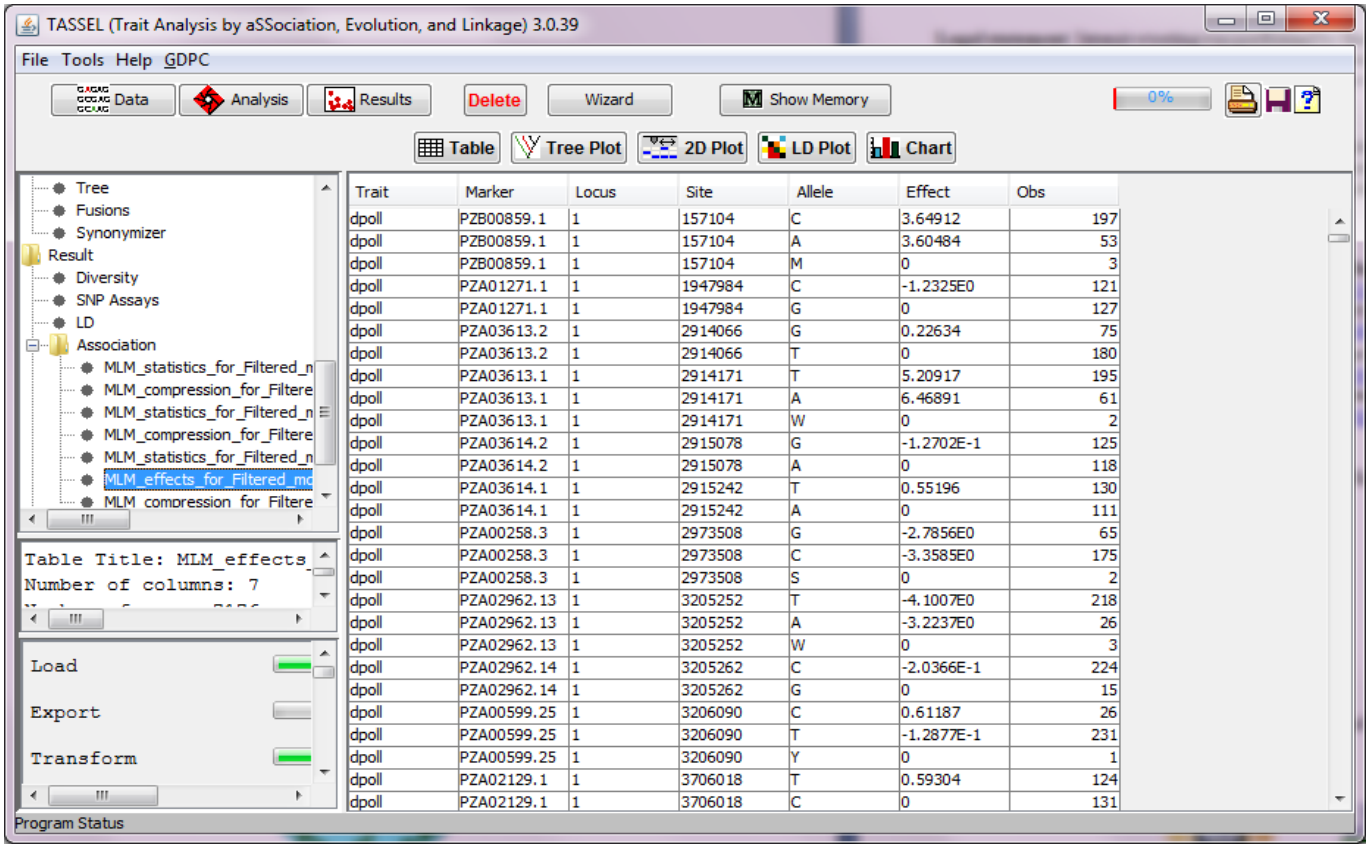
**Note**: When two or more individuals are clustered into one group, the variance component for the random effect is not equivalent to the one without compression. Consequently, the heritability derived should not be interpreted as the individual based heritability.

To perform a Genome-Wide Association Study (GWAS) on the 3093 SNPs, we need to create a new joint data set containing the filtered phenotype, population structure, and the genome–wide genotype. Highlight the new joint file ar the kinship data and click the **MLM** button. Choose the default options on the MLM option dialog. The analysis will take a minute or two. The output report labeled "MLM_compression" indicates that 259 lines were used in the analysis. With 74 groups, the statistics from the best are as graphed below.




The strongest associated SNP is at 193565357 bp on chromosome 3. The P value is $1.3027 \times 10^{-4}$. The threshold is $3.2331 \times 10^{-5}$ at significant level of 1% after Bonferroni multiple test correction (0.01/3093). The association was not significant. As illustrated below, the output labeled "GLM_Allele_Estimates" shows the marker effects assigned t genotypes for each SNP (The GLM is also the same). For example, the first SNP at 157104 bp on chromosome 1 had three genotypes (AA, CC and AC) coded as A, C, and M based on the IUPAC code, see Appendix (Nucleotide Codes).




| Trait | Marker | Locus | Site | Allele | Effect | Obs |
|---|---|---|---|---|---|---|
| dpoll | PZB00859.1 | 1 | 157104 | C | 3.64912 | 197 |
| dpoll | PZB00859.1 | 1 | 157104 | A | 3.60484 | 53 |
| dpoll | PZB00859.1 | 1 | 157104 | M | 0 | 3 |
| dpoll | PZA01271.1 | 1 | 1947984 | C | -1.2325E0 | 121 |
| dpoll | PZA01271.1 | 1 | 1947984 | G | 0 | 127 |
| dpoll | PZA03613.2 | 1 | 2914066 | G | 0.22634 | 75 |
| dpoll | PZA03613.2 | 1 | 2914066 | T | 0 | 180 |
| dpoll | PZA03613.1 | 1 | 2914171 | T | 5.20917 | 195 |
| dpoll | PZA03613.1 | 1 | 2914171 | A | 6.46891 | 61 |
| dpoll | PZA03613.1 | 1 | 2914171 | W | 0 | 2 |
| dpoll | PZA03614.2 | 1 | 2915078 | G | -1.2702E-1 | 125 |
| dpoll | PZA03614.2 | 1 | 2915078 | A | 0 | 118 |
| dpoll | PZA03614.1 | 1 | 2915242 | T | 0.55196 | 130 |
| dpoll | PZA03614.1 | 1 | 2915242 | A | 0 | 111 |
| dpoll | PZA00258.3 | 1 | 2973508 | G | -2.7856E0 | 65 |
| dpoll | PZA00258.3 | 1 | 2973508 | C | -3.3585E0 | 175 |
| dpoll | PZA00258.3 | 1 | 2973508 | S | 0 | 2 |
| dpoll | PZA02962.13 | 1 | 3205252 | T | -4.1007E0 | 218 |
| dpoll | PZA02962.13 | 1 | 3205252 | A | -3.2237E0 | 26 |
| dpoll | PZA02962.13 | 1 | 3205252 | W | 0 | 3 |
| dpoll | PZA02962.14 | 1 | 3205262 | C | -2.0366E-1 | 224 |
| dpoll | PZA02962.14 | 1 | 3205262 | G | 0 | 15 |
| dpoll | PZA00599.25 | 1 | 3206090 | C | 0.61187 | 26 |
| dpoll | PZA00599.25 | 1 | 3206090 | T | -1.2877E-1 | 231 |
| dpoll | PZA00599.25 | 1 | 3206090 | Y | 0 | 1 |
| dpoll | PZA02129.1 | 1 | 3706018 | T | 0.59304 | 124 |
| dpoll | PZA02129.1 | 1 | 3706018 | C | 0 | 131 |

# 11 Appendix

## *11.1Nucleotide Codes (Derived from IUPAC)*

| Code | Meaning |
|---|---|
| **A** | **A:A** |
| **C** | **C:C** |
| **G** | **G:G** |
| **T** | **T:T** |
| **R** | **A:G** |
| **Y** | **C:T** |
| **S** | **C:G** |
| **W** | **A:T** |
| **K** | **G:T** |
| **M** | **A:C** |
| **+** | **+:+ (insertion homozygous)** |
| **0** | **+:-** |
| **-** | **-:- (deletion homozygous)** |
| **N** | **Unknown** |

## *11.2TASSEL Tutorial Data sets*

http://www.maizegenetics.net/tassel/docs/TASSELTutorialData3.zip

| File name | Type | Format |
|---|---|---|
| d8_sequence.phy | Genotype | Phylip Alignment |
| mdp_genotype.hmp.txt | Genotype | Hapmap Alignment |
| mdp_genotype.plk.ped<br>mdp_genotype.plk.map | Genotype | Plink Alignment |
| mdp_kinship.txt | Kinship | Numerical square matrix |
| mdp_population_structure.txt | Population structure | Numerical trait data |
| mdp_traits.txt | Phenotype | Numerical trait data |

File #1 is the sequence of dwarf8 gene with 2466 sites on 91 maize inbred lines. The data was described by the paper on the association between Dwarf8 and flowering time[26].

File #2-6 are 3093 SNPs on 281 maize association inbred lines. The data was presented in three formats (Hapmap, Plink and Flapjack). The data was created by the PANZEA project funded by NSF. Details of the data can be found at http://www.panzea.org.

File #5 and 6 are in pair for the format of Plink.

File #7 is kinship created by Yu et al.[9].

File #8 is population structure of 282 maize inbred line[27].

File #9 is phenotype on three traits, including flowering time, on 282 maize inbred lines[9].

## *11.3Frequently Asked Questions*

1. **What do I do if TASSEL misbehaves?**
   TASSEL is an open source software project hosted on SourceForge and has a bug tracking list at http://sf.net/projects/tassel where you can notify the developer community of problems. In order for a bug to be fixed, we must be able to replicate the problem. Thus, it is important to document the steps that were taken that produced the error. If the data you are working with is not too sensitive, please include the files which were used in the faulty procedure. If you would rather not post your data file on SourceForge, you may email it to one of the software developers.

2. **Where do I turn for more information?**
   If you are having difficulty with a certain aspect of TASSEL, you can either email one of the software developers listed at www.maizegenetics.net or you may check the TASSEL forum on SourceForge http://sf.net/projects/tassel), as another user may have already addressed a similar question. There is also a TASSEL discussion group at http://groups.google.com/group/tassel.

3. **How do I join the fun: TASSEL on SourceForge?**
   TASSEL is an open source project distributed under the GNU general public license. This means that the source code is available and the user is free to modify the code to suit their particular needs. We welcome input from developers and those who wish to become involved in the improvement of this software. The project is hosted on SourceForge (http://sf.net/projects/tassel), thereby allowing anyone to access the most recent changes to the code. This setup makes it convenient for anyone to add special functionality to TASSEL if they so desire. It also serves as a good platform for anyone who wishes to become involved in a bioinformatics software development project.

4. **When I click on the most current version of TASSEL web start, a previous version appears. What should I do?**
   The previous version of TASSEL web start was cached in your machine. To replace it with the most current version, click the Start button in Windows, followed by Run. Type **javaws** and then click OK. In the window that opens, keep the most current version of TASSEL and delete the rest.

5. **What should I substitute for missing values in TASSEL?**
   For numerical data in version 3 format, use NA or NaN. For numerical data in version 2 format, use "-999" for missing values. For SNP data, use "N". Kinship does not allow missing values.

6. **Is it possible to change data names in the Data Tree?**
   Yes. Click on the desired data name in the Data Tree, wait for one second, and then click it again or immediately hit the F2 key. Rename the data set and then hit Enter to save the change.

7. **How can I create a TASSEL icon on desktop?**
   Click "Start" on Microsoft Windows and select "Control Panel", then double click Java to show "java Control Panel". In "Temporary Internet Files" section, click "View" button show "Java Cache Viewer". Move mouse over TASSEL application and click right button and select "Install Shortcuts".

8. **Why do I get empty squares in MLM association analysis?**
   The empty square means null information. The major reasons include non-convergence in the estimation of variance components or that the statistic in question was not calculated. For example, marker F, p, and $R^2$ are not calculated

when no marker is included in the model.

9. **Why should I exclude one column of the population structure?**

For some methods of calculating population structure, such as the software STRUCTURE, the population proportions sum to one. This produces linear dependence between the population co-variates. While the algorithm used by GLM tolerates that dependency, MLM will fail because the design matrix will not be invertible Excluding one column eliminates linear dependence between columns. Using PC axes to represent population structure does not result in linear dependency because all PC columns are guaranteed to be independent.

**10. Can kinship replace population structure?**

Sometimes. For some traits and populations, the K-only model may be as good as or better than the Q+K model. For others, Q+K may be superior. The Q-only model is not as effective for controlling population structure as the alternatives. Unfortunately, no general guidelines exist for predicting which model will perform best. As a result, a investigator may wish to fit all three models and compare the results. If eliminating false positives is very importa then it may make sense to accept the most conservative model. However, if the objective is to identify candidates for further study and the cost of following up on a false lead is low, the most liberal model may be preferred.

**11. Why do TASSEL and SPAGeDi give different kinship estimates?**

First, many algorithms exist to calculate kinship and their estimates will differ from one another. Secondly, th algorithm in TASSEL treats each genotype as a haplotype. It is not recommended that TASSEL be used to generate a kinship matrix from heterozygous genotype. In the near future, the TASSEL kinship algorithm will be modified handle heterozygous diploids.

**12. Can I get Marker R square using SAS Proc Mixed or TASSEL MLM?**

SAS Proc Mixed does not produce an $R^2$ statistic. MLM in TASSEL does. The user manual describes how it is calculated.

**13. Does MLM find more associations than GLM?**

Sometimes. MLM has higher statistical power than GLM and may detect more true associations.. When the tested genetic markers are confounded with kinship structure , GLM does not correct for that as effectively as MLM and may produce more false positives

**14. Do I need multiple test correction for the p value from Tassel?**

Yes.

**15. Can TASSEL handle diploid genotype data?**

While TASSEL accepts most common sequence alignment formats which handle polyploid genotype data including haploid and diploid, some analyses are not appropriate for heterozygous data. GLM or MLM fit SNPs one at a time, treating each distinct genotype as a separate class. This has the effect of fitting an additive plus dominance model. Separating the two effects is under consideration. Because handling heterozygotes as a third marker class is not appropriate for kinship or LD those analyses should not be used for that type of data at the present time. Work to improve handling heterozygotes is ongoing.

**16. How to cite TASSEL?**

The paper that describes TASSEL[1] as a software package and the papers that introduce specific methods implemented in TASSEL should be cited as appropriate, such as the unified ("Q+K") approach, EMMA, compression of mixed linear model and P3D. For example,:

A. Linkage disequilibrium (D', $R^2$ and P value) were calculated by TASSEL[1].
B. Association analyses were performed with the mixed linear model approach[9] implemented by TASSEL[1].
C. GWAS was performed with the compressed mixed linear model approach[4,9] carried by TASSEL[1] which also implemented the EMMA[3] and P3D[4] algorithms to reduce computing time.

## *REFERENCES*


1. Bradbury, P.J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635 (2007).
2. Zhang, Z., Buckler, E.S., Casstevens, T.M. & Bradbury, P.J. Software engineering the mixed model for genome-wide association studies on large samples. *Brief Bioinform* **10**, 664-75 (2009).
3. Kang, H.M. et al. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178**, 1709-1723 (2008).
4. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355-60 (2010).
5. Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
6. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28**, 286-289 (2001).
7. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170-181 (2000).
8. Zhao, K. et al. An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
9. Yu, J.M. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-208 (2006).
11. Ware, D. et al. Gramene: a resource for comparative grass genomics. *Nucleic Acids Research* **30**, 103-105 (2002).
12. Ware, D.H. et al. Gramene, a tool for grass Genomics. *Plant Physiology* **130**, 1606-1613 (2002).
13. Jaiswal, P. et al. Gramene: development and integration of trait and gene ontologies for rice. *Comparative and Functional Genomics* **3**, 132-136 (2002).
14. Yamazaki, Y. & Jaiswal, P. Biological ontologies in rice databases. An introduction to the activities in gramene and oryzabase. *Plant and Cell*

*Physiology* **46**, 63-68 (2005).
15. Zhao, W. et al. Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Research* **34**, D752-D757 (2006).
16. Canaran, P., Stein, L. & Ware, D. Look-Align: an interactive web-based multiple sequence alignment viewer with polymorphism analysis support. *Bioinformatics* **22**, 885-886 (2006).
17. Du, C.G., Buckler, E. & Muse, S. Development of a maize molecular evolutionary genomic database. *Comparative and Functional Genomics* **4**, 246-249 (2003).
18. SAS, I.I. SAS. Statistical Analysis Software for Windows, 9.0 ed. *Cary, NC. USA.* ( 2002.).
19. Hardy, O.J. & Vekemans, X. SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* **2**, 618-620 (2002).
20. Cover, T. & Hart, P. Nearest neighbor pattern classification. *Proc IEEE Trans Inform Theory* **13**(1967).
21. Weir. Genetic Data Analysis II. *Sunderland, MA.* (1996).
22. Farnir, F. et al. Extensive genome-wide linkage disequilibrium in cattle. *Genome Res* **10**, 220-7 (2000).
23. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**, 423-447 (1975).
24. Kang, H.M. et al. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).
25. Laird, N.M. & Ware, J.H. Random-Effects Models for Longitudinal Data. *Biometrics* **38**, 963-974 (1982).
26. Thornsberry, J.M. et al. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* **28**, 286-9 (2001).
27. Flint-Garcia, S.A. et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* **44**, 1054-64 (2005).
28. Anderson, M.J. & Ter Braak, C.J.F. Permutations tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73, 85-113 (2003)